Wage and its dependence on other factors

Tran To Bang Trinh

Contents

1 About the dataset

1.1 Description

- This dataset is `cps4c_small.csv`, which is provided in "Principles of Econometrics" by Carter Hill, William Griffiths, and Guay Lim, 4th edition, Wiley, into R.
- The dataset contains 1000 non-duplicated rows and 12 columns, including: *wage*, *educ*, *exper*, *hrswk*, *married*, *female*, *metro*, *midwest*, *south*, *west*, *black*, and *asian*.

1.2 Structure of the dataset

Detail of all the columns (attributes) of the dataset:

No	Variables	Definition	Unit	Type
1	wage	Earnings per hour	dollars	numeric
2	educ	Years of education	years	numeric
3	exper	Years of experience (post education)	years	numeric
4	hrswk	Weekly working hours	hours	numeric
5	married	Martial status, 1 if married, 0 otherwise		categorical
6	female	Gender, 1 if female, 0 otherwise		categorical
7	metro	Living in metropolitan area, 1 if yes, 0 otherwise		categorical
8	midwest	Living area, 1 if in midwest, 0 otherwise		categorical
9	south	Living area, 1 if in south, 0 otherwise		categorical
10	west	Living area, 1 if in west, 0 otherwise		categorical
11	black	Race, 1 if black people, 0 otherwise		categorical
12	asian	Race, 1 if asian people, 0 otherwise		categorical

1.3 Data pre-processing

- Load the dataset `cps4c_small.csv` into *data* variable.
- Save data for further usage.
- Use `attach` function to retrieve each variable by calling its name.
- Use `ncol` function to count the number of variables in the dataset.
- Use `nrow` function to count the number of tuples in the dataset.

Code:

Input helpful libraries

```
library(dplyr)
library(ggplot2)
library(viridis)
library(scales)
library(cowplot)
library(showtext)
library(ggrepel)
library(tidyverse)
library(GGally)
```

Set up the default font for our graphs.

```
font_add_google("Raleway", family = "Title")
font_add_google("Bitter", family = "Axis")
font_add_google("Kameron", family = "Text")
showtext_auto()
```

Set up working directory and load the data set

```
setwd("C:/Users/Rachel/Documents/R/Project")
getwd()
```

```
## [1] "C:/Users/Rachel/Documents/R/Project"
```

```
data<-read.csv("cps4c_small.csv", header=TRUE, stringsAsFactors = FALSE)
```

Preprocess the data to get the frequencies of qualitative data

```
data$married <- factor(data$married)
levels(data$married) <- c("Single", "Married")
data$female <- factor(data$female)
levels(data$female) <- c("Male", "Female")
data$metro <- factor(data$metro)
levels(data$metro) <- c("Outskirts", "Metro")
data$midwest <- factor(data$midwest)
data$south <- factor(data$south)
data$west <- factor(data$west)
data$black <- factor(data$black)
data$asian <- factor(data$asian)
```

```
attach(data)
```

Back up the data

```
save(data, file = 'wage.rda')
```

The number of rows of the dataset

```
nrow(data)
```

```
## [1] 1000
```

The number of columns of the dataset

```
ncol(data)
```

```
## [1] 12
```

Take a look of the data:

- By using **head** function, we can display the sample data (6 first rows) of the dataset

```
head(data)

##    wage educ exper hrswk married female     metro midwest south west black
## 1 18.70   16    39    37 Married Female     Metro       0     1    0     0
## 2 11.50   12    16    62  Single   Male Outskirts       1     0    0     0
## 3 15.04   16    13    40 Married   Male     Metro       0     0    1     1
## 4 25.95   14    11    40  Single Female     Metro       0     1    0     1
## 5 24.03   12    51    40 Married   Male     Metro       0     0    0     0
## 6 20.00   12    30    40 Married   Male     Metro       0     0    0     0
##   asian
## 1     0
## 2     0
## 3     0
## 4     0
## 5     0
## 6     0
```

- By using **str** function

```
str(data)

## 'data.frame':    1000 obs. of  12 variables:
##  $ wage   : num  18.7 11.5 15 25.9 24 ...
##  $ educ   : int  16 12 16 14 12 12 16 12 16 13 ...
##  $ exper  : int  39 16 13 11 51 30 30 34 31 31 ...
##  $ hrswk  : int  37 62 40 40 40 40 50 60 45 40 ...
##  $ married: Factor w/ 2 levels "Single","Married": 2 1 2 1 2 2 2 2 2 1 ...
##  $ female : Factor w/ 2 levels "Male","Female": 2 1 1 2 1 1 1 1 1 2 ...
##  $ metro  : Factor w/ 2 levels "Outskirts","Metro": 2 1 2 2 2 2 2 2 2 2 ...
##  $ midwest: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 2 1 ...
##  $ south  : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
##  $ west   : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 1 1 1 ...
##  $ black  : Factor w/ 2 levels "0","1": 1 1 2 2 1 1 1 1 1 1 ...
##  $ asian  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

- By using **summary** function, it will give us:
- Qualitative data: The frequencies of each type.
- Quantitative data: 5 point summary (Min, First Quantiles, Median, Mean, Third Quantile, Max).

```
summary(data)

##       wage             educ          exper           hrswk          married
##  Min.   : 1.97   Min.   : 0.0   Min.   : 2.00   Min.   : 0.00   Single :419
##  1st Qu.:11.25   1st Qu.:12.0   1st Qu.:16.00   1st Qu.:40.00   Married:581
##  Median :17.30   Median :13.0   Median :27.00   Median :40.00
##  Mean   :20.62   Mean   :13.8   Mean   :26.51   Mean   :39.95
##     female          metro     midwest south   west    black   asian
##  Male  :486   Outskirts:220   0:760   0:704   0:760   0:888   0:957
##  Female:514   Metro    :780   1:240   1:296   1:240   1:112   1: 43
##
##
##  [ reached getOption("max.print") -- omitted 2 rows ]
```

2 Descriptive Statistics

2.1 Quantitative variables

2.1.1 Attribute *wage*

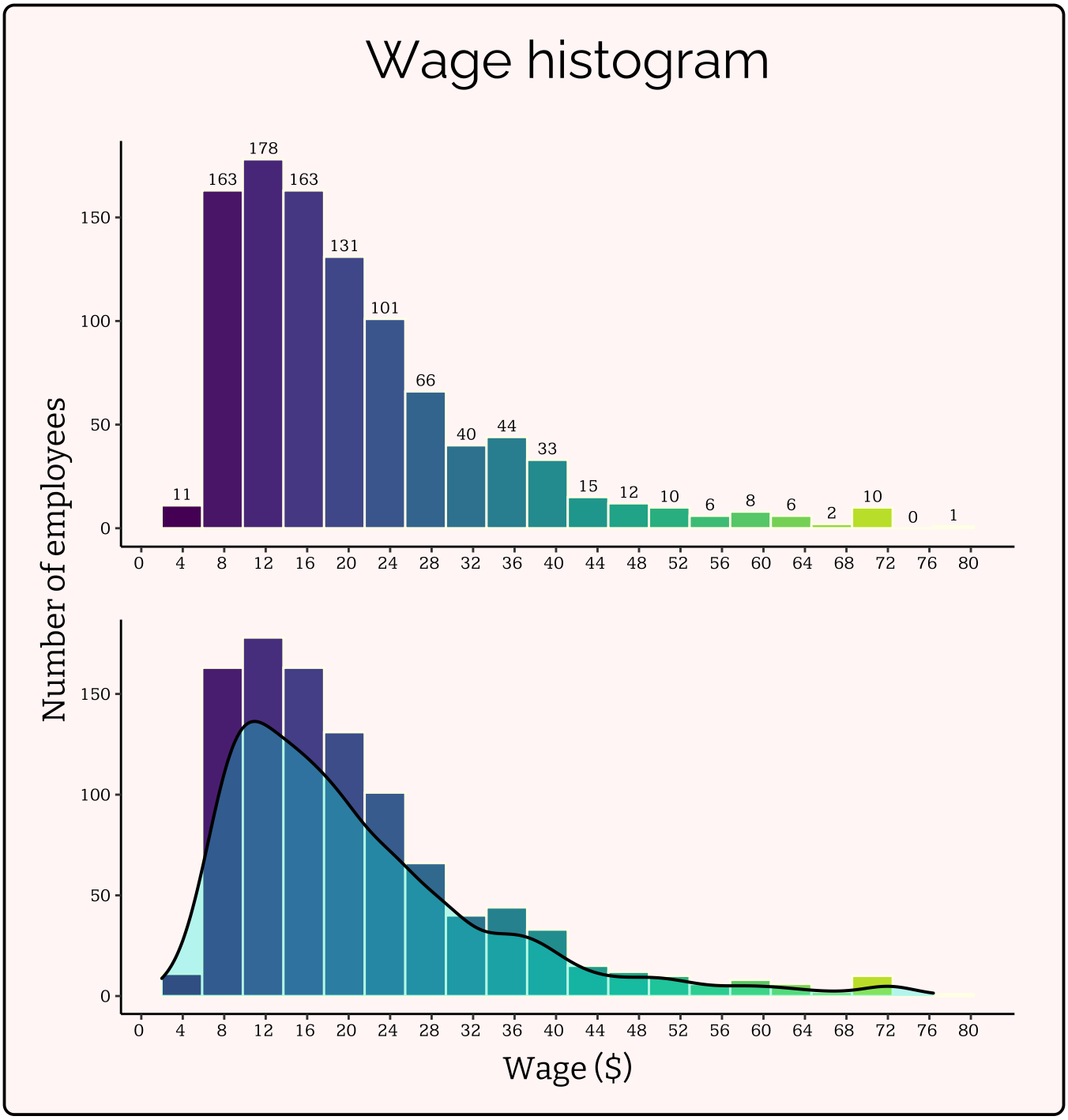
- Put the *wage* data into dataframe

```
df <- data.frame(x = wage)
summary(df)

##        x
##  Min.   : 1.97
##  1st Qu.:11.25
##  Median :17.30
##  Mean   :20.62
##  3rd Qu.:25.63
##  Max.   :76.39
```

- Plot *wage* as histogram to view its distribution

```
a <- ggplot(df, aes(x = x, fill=..x.., y=..count..)) +
  geom_histogram(col="#ffffe6", bins=20) +
  stat_bin(aes(y=..count.., label=..count..), geom="text", vjust= -0.5, size=3,
           bins = 20, family="Text") +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 80, 4)) +
  theme_classic() +
  labs(title = "Wage histogram", x = "", y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        plot.title = element_text(hjust = 0.5, family = "Title", size=25,
                                  margin=margin(0,0,30,0)),
        axis.title = element_text(family = "Axis", size=15),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

b <- ggplot(df, aes(x = x, fill=..x.., y=..count..)) +
  geom_histogram(col="#ffffe6", bins=20) +
  geom_density(aes(y = ..density.. * nrow(df) * 3), lwd = 0.7, colour = "#000000",
               fill = "#00F2DE", alpha = 0.3) +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 80, 4)) +
  theme_classic() +
  labs(x = "Wage ($)",
```

```
       y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

plot_grid(a, b, align="v", nrow=2, rel_heights = c(1.2, 1)) +

  draw_label("Number of employees", x = 0, y=0.5, vjust= 1.1, angle=90,
             fontfamily = "Axis", size=15)
```

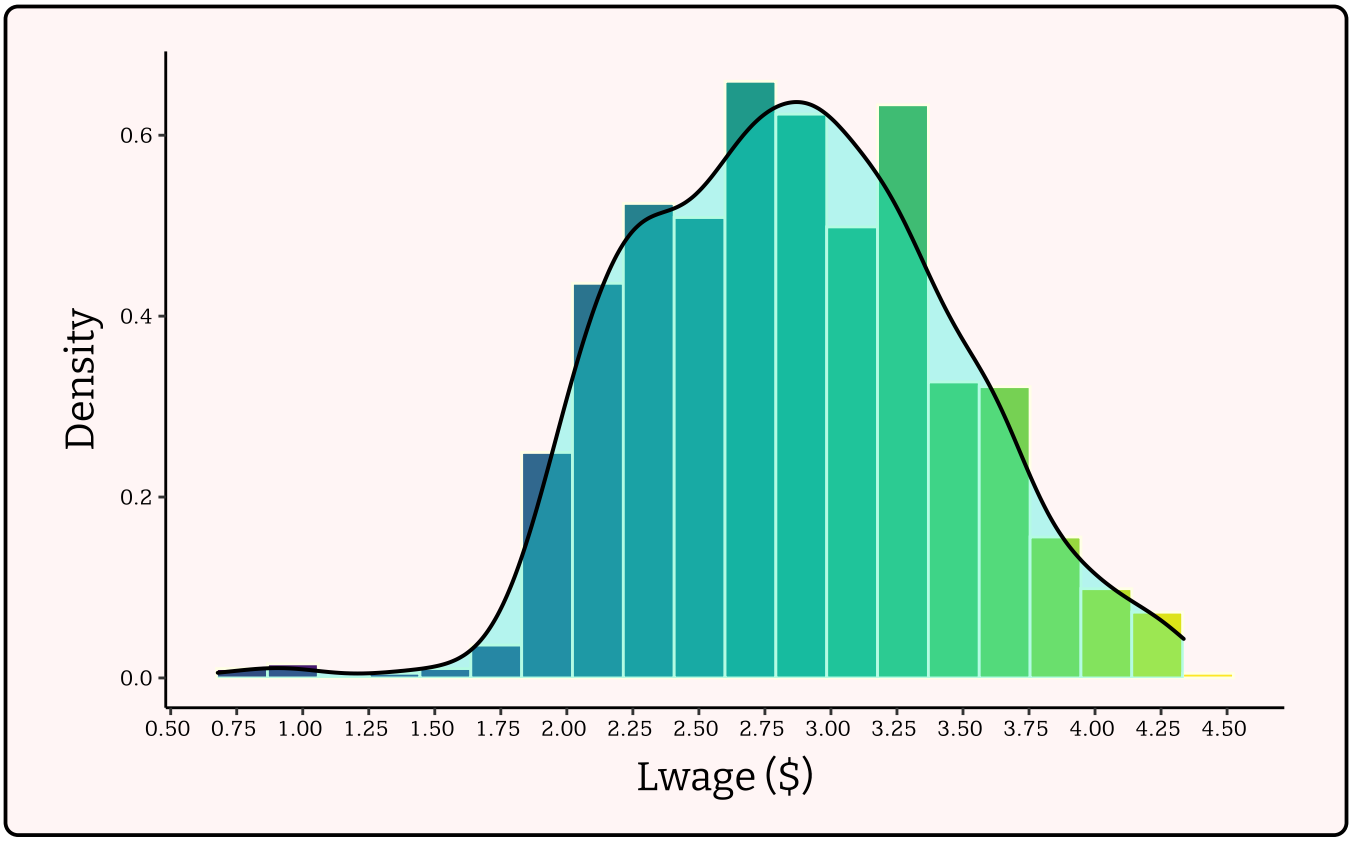


- **Remark:**
  - The distribution of wage is right-skewed, mostly from 8 to 40, it implies that the earnings of the majority is in the range of \$8 and \$40 per hour.
  - The mode of wage is at \$12 with 178 people, the largest proportion, having that same earning.
  - Adding the density curve, we can see that the higher the wage, the fewer people earn that amount of money. We also notice that the wage's distribution has quite similar form to Chi-squared (χ^2) distribution.
- Because *wage* does not have the normal distribution, calculate $ln(wage)$ as *lwage* to normalize it.

```
lwage = log(wage)
data$lwage = lwage

df <- data.frame(x = lwage)
summary(df)

##        x
##  Min.   :0.678
##  1st Qu.:2.420
##  Median :2.851
##  Mean   :2.857
##  3rd Qu.:3.244
##  Max.   :4.336
```

- Plot histogram to view its normal distribution form.

```
ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(aes(y = ..density..), col="#ffffe6", bins=20) +
  geom_density(lwd = 0.7, colour = "#000000",
               fill = "#00F2DE", alpha = 0.3) +

  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 5, 0.25)) +
  theme_classic() +
  labs(x = "Lwage ($)",
       y = "Density") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```



2.1.2 Attribute *educ*

- Put the *educ* data into dataframe

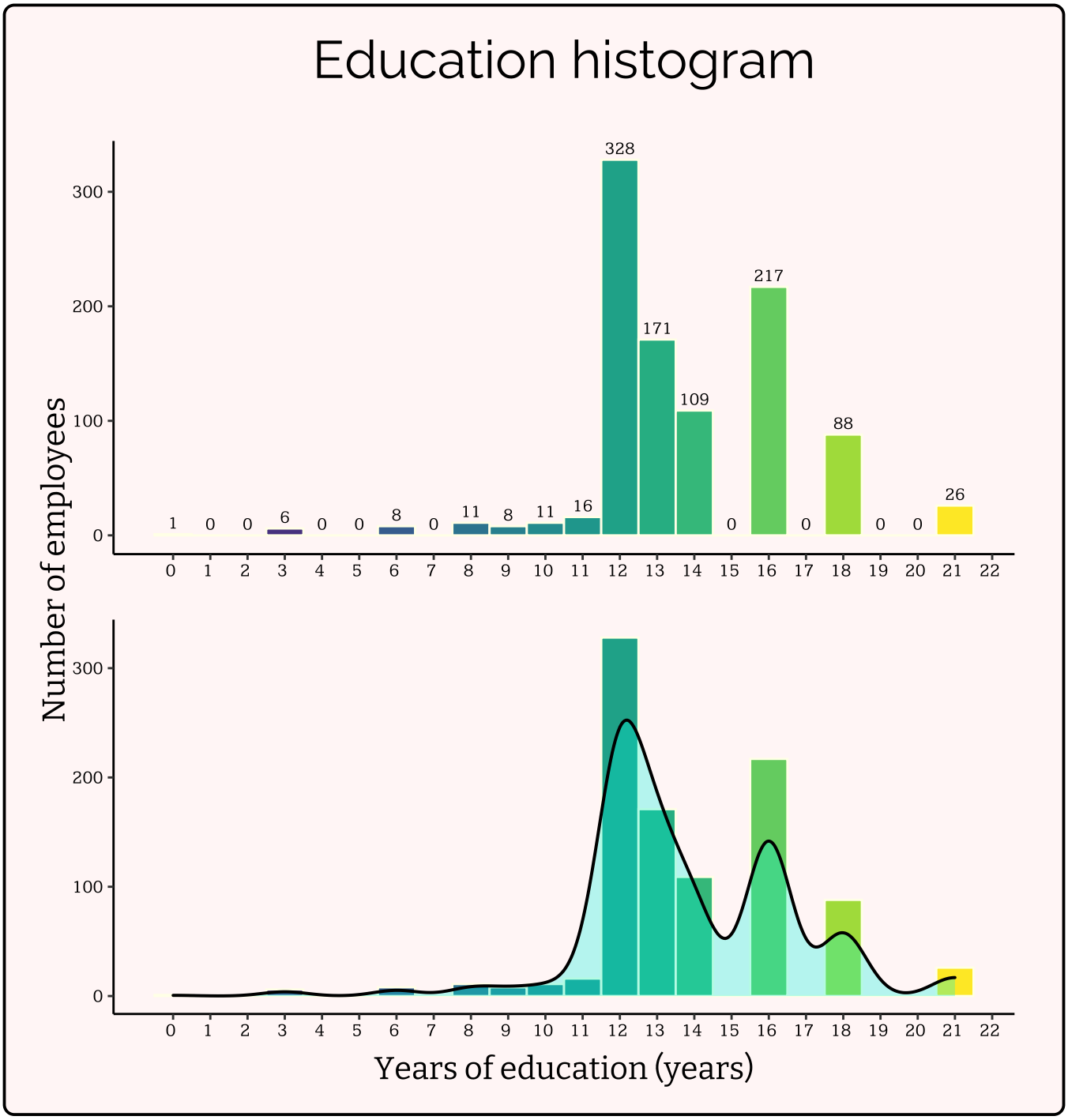
```
df <- data.frame(x = educ)
summary(df)
```

```
##        x
##  Min.   : 0.0
##  1st Qu.:12.0
##  Median :13.0
##  Mean   :13.8
##  3rd Qu.:16.0
##  Max.   :21.0
```

- Plot *educ* data as histogram to view its distribution

```
a <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=22) +
  stat_bin(aes(y=..count.., label=..count..), geom="text", vjust= -0.5, size=3,
           bins = 22, family="Text") +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 22, 1)) +
  theme_classic() +
```

```
  labs(title = "Education histogram",
       x = "",
       y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        plot.title = element_text(hjust = 0.5, family = "Title", size=25,
                                  margin=margin(0,0,30,0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

b <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=22) +
  geom_density(aes(y = ..density.. * nrow(df)), lwd = 0.7, colour = "#000000",
               fill = "#00F2DE", alpha = 0.3) +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 22, 1)) +
  theme_classic() +
  labs(x = "Years of education (years)",
       y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0), family = "Axis",
                                    size=15),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

plot_grid(a, b, rel_heights = c(1.2, 1), align="v", nrow=2) +
  draw_label("Number of employees", x = 0, y=0.5, vjust= 1.1, angle=90,
             fontfamily = "Axis", size=15)
```



- **Remark:**
  - The distribution of education is quite almost bimodal having two distinct peaks at 12, who was graduated from high school, and 16, who got Bachelor's degree.
  - High school completed is the most common education level of the employee, with 328 people. The second peak is at University graduation with 217 people. Other milestones that the majority of data occupies at is 13, 14, 18 and 21 in the descending order of quantity.
  - Adding the density curve, we can see that the education's distribution reach its peak at 12, and oscillates damply to the end.

2.1.3 Attribute *exper*

- Put the *exper* data into dataframe

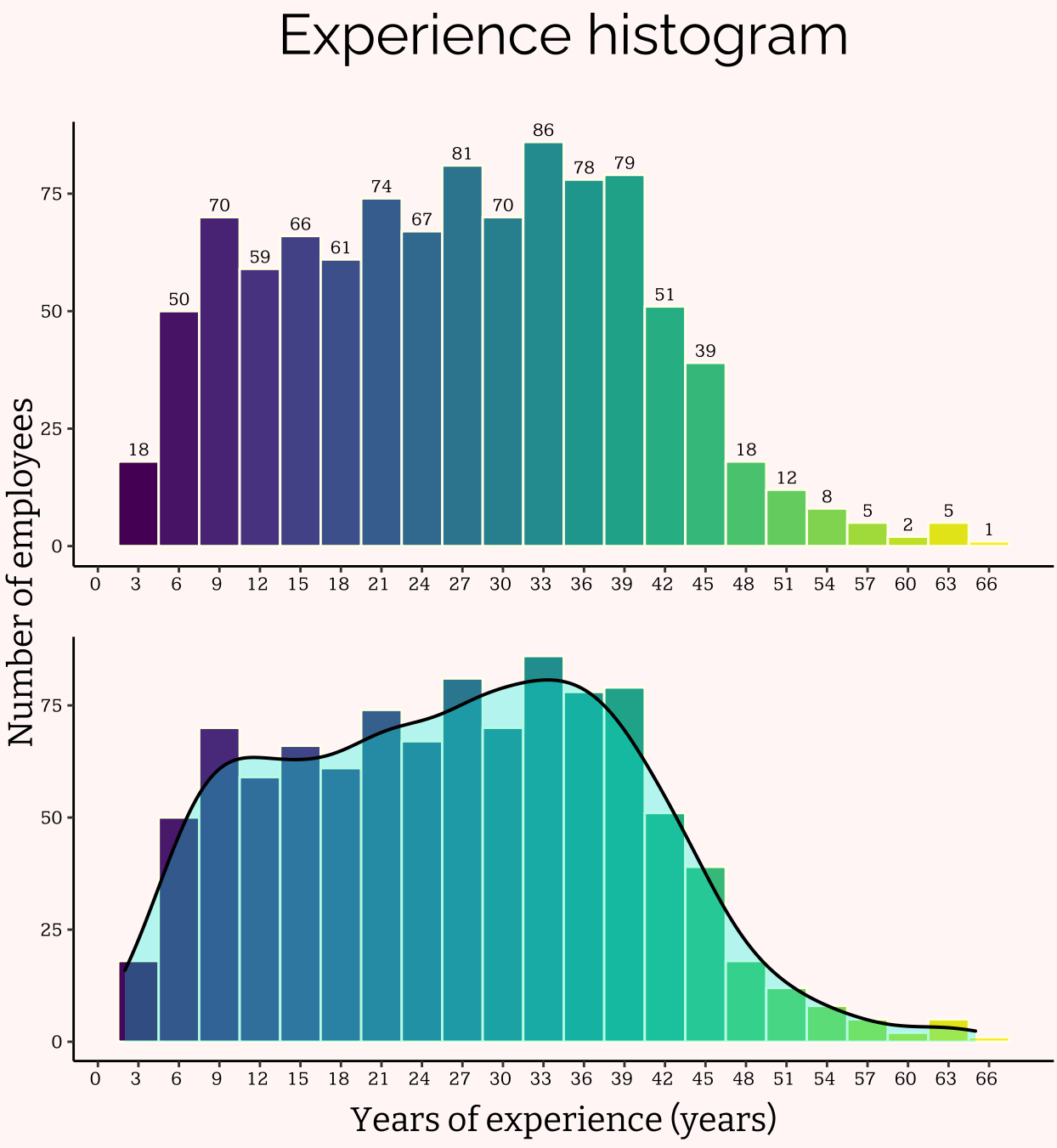
```
df <- data.frame(x = exper)
summary(df)
```

```
##        x
##  Min.   : 2.00
##  1st Qu.:16.00
##  Median :27.00
##  Mean   :26.51
##  3rd Qu.:36.00
##  Max.   :65.00
```

- Plot *exper* data as histogram to view its distribution

```
a <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=22) +
  stat_bin(aes(y=..count.., label=..count..), geom="text", vjust= -0.5, size=3,
           bins = 22, family="Text") +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 66, 3)) +
  theme_classic() +
  labs(title="Experience histogram", x="", y="") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        plot.title = element_text(hjust = 0.5, family = "Title", size=25,
                                  margin=margin(0,0,30,0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

b <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=22) +
  geom_density(aes(y = ..density.. * nrow(df) * 3), lwd = 0.7, colour = "#000000",
               fill = "#00F2DE", alpha = 0.3) +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 66, 3)) +
  theme_classic() +
  labs(x="Years of experience (years)", y="") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
```

```
        axis.text = element_text(family = "Text", color="#000000"),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0), family = "Axis",
                                    size=15),
        text = element_text(family = "Text", color="#000000"))

plot_grid(a, b, align="v", nrow=2, rel_heights = c(1.2, 1)) +
  draw_label("Number of employees", x = 0, y=0.5, vjust= 1.1, angle=90,
             fontfamily = "Axis", size=15)
```



- **Remark:**
  - The years of experience of the workers are diverse, but mostly under 48 years.

- It's obvious that the distribution of the experience is almost uniform from 9 to 39 years, the experience level most of the employees have with more than 60 people.
- Adding the density curve, we can see that after the uniform distribution, the graph tends to go downward as the experience years grows due to the retirement, therefore, the number of enployees that still work at that age is not too many.

2.1.4 Attribute *hrswk*

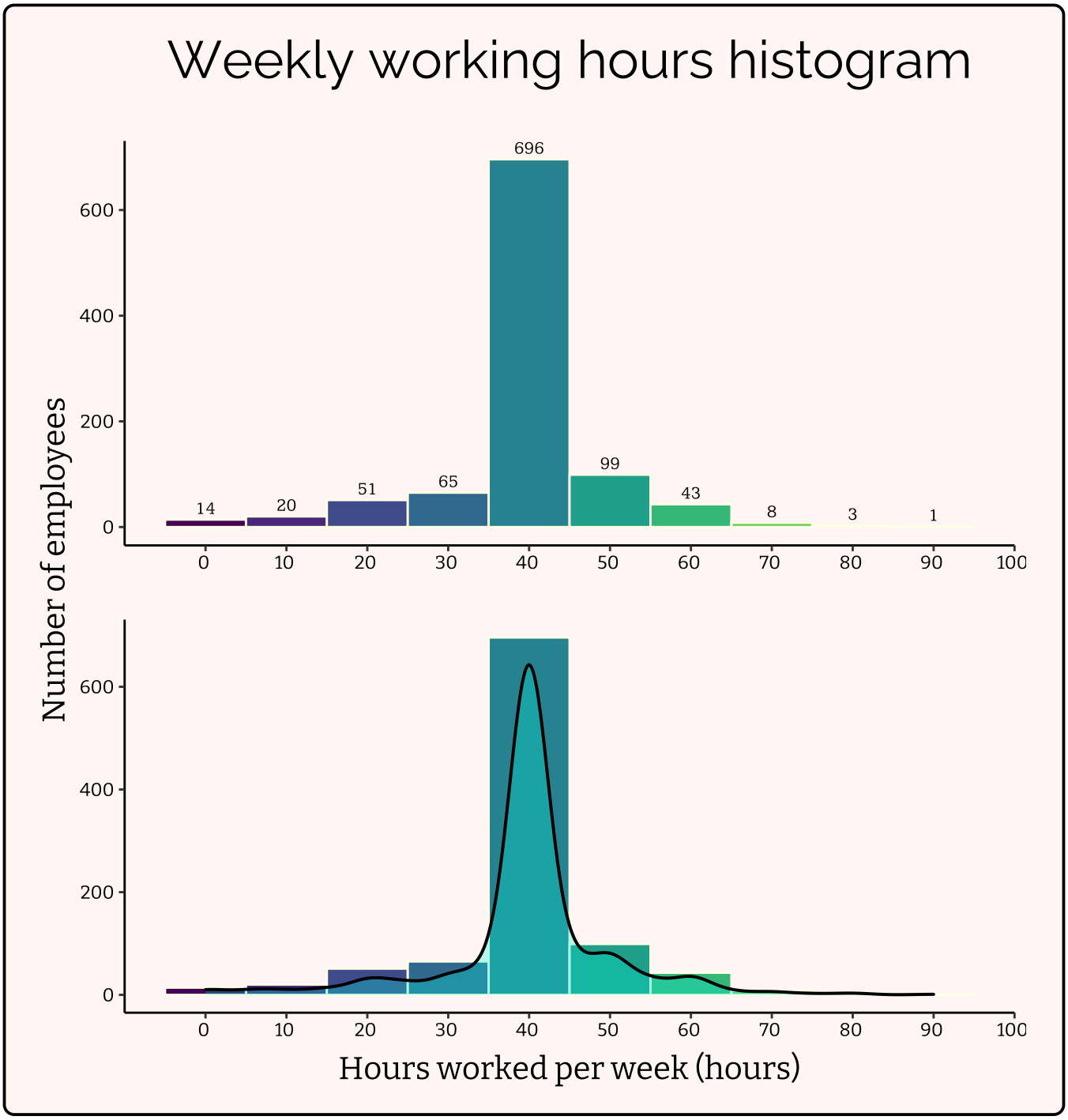
- Put *hrswk* data into dataframe

```
df <- data.frame(x = hrswk)
summary(df)

##        x
##  Min.   : 0.00
##  1st Qu.:40.00
##  Median :40.00
##  Mean   :39.95
##  3rd Qu.:40.00
##  Max.   :90.00
```

- Plot *hrswk* data as histogram to view its distribution

```
a <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=10) +
  stat_bin(aes(y=..count.., label=..count..), geom="text", vjust= -0.5, size=3,
           bins = 10, family="Text") +
  scale_fill_viridis() +

  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 100, 10)) +
  theme_classic() +
  labs(title = "Weekly working hours histogram",
       x = "",
       y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        plot.title = element_text(hjust = 0.5, family = "Title", size=25,
                                  margin=margin(0,0,30,0)),
        axis.title = element_text(family = "Axis", size=15),
        axis.text = element_text(family = "Axis", color="black"))


b <- ggplot(df, aes(x = x, fill=..x..)) +
  geom_histogram(col="#ffffe6", bins=10) +
  geom_density(aes(y = ..density.. * nrow(data) * 6), lwd = 0.7, colour = "#000000",
               fill = "#00F2DE", alpha = 0.3) +
  scale_fill_viridis() +
```

```
  scale_x_continuous(labels = unit_format(unit=""),
                     breaks=seq(0, 100, 10)) +
  theme_classic() +
  labs(x = "Hours worked per week (hours)",
       y = "") +
  theme(legend.position="none",
        plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0), family = "Axis",
                                    size=15),
        axis.text = element_text(family = "Axis", color="black"))

plot_grid(a, b, align="v", nrow=2, rel_heights = c(1.2, 1)) +
  draw_label("Number of employees", x = 0, y=0.5, vjust= 1.1, angle=90,
             fontfamily = "Axis", size=15)
```



- **Remark:**
  - The distribution of *hrswk* has an outstanding peak at about 40 hours/week, which is the standard of full-time employment working hours.
  - Nearly 70% people in the data work as an ordinary full-time employee, which are 696 people in total. There are also some popular working hours from 0 to 60, but occupies not too many comparing to the peak.
  - Viewing the density curve, we can see that the distribution of *hrswk* data resembles normal distribution.

2.2 Qualitative variables

2.2.1 Attribute *married*

- Count the data and calculate its proportion

```
married_table <- table(married)
married_table

## married
##  Single Married
##     419     581
```

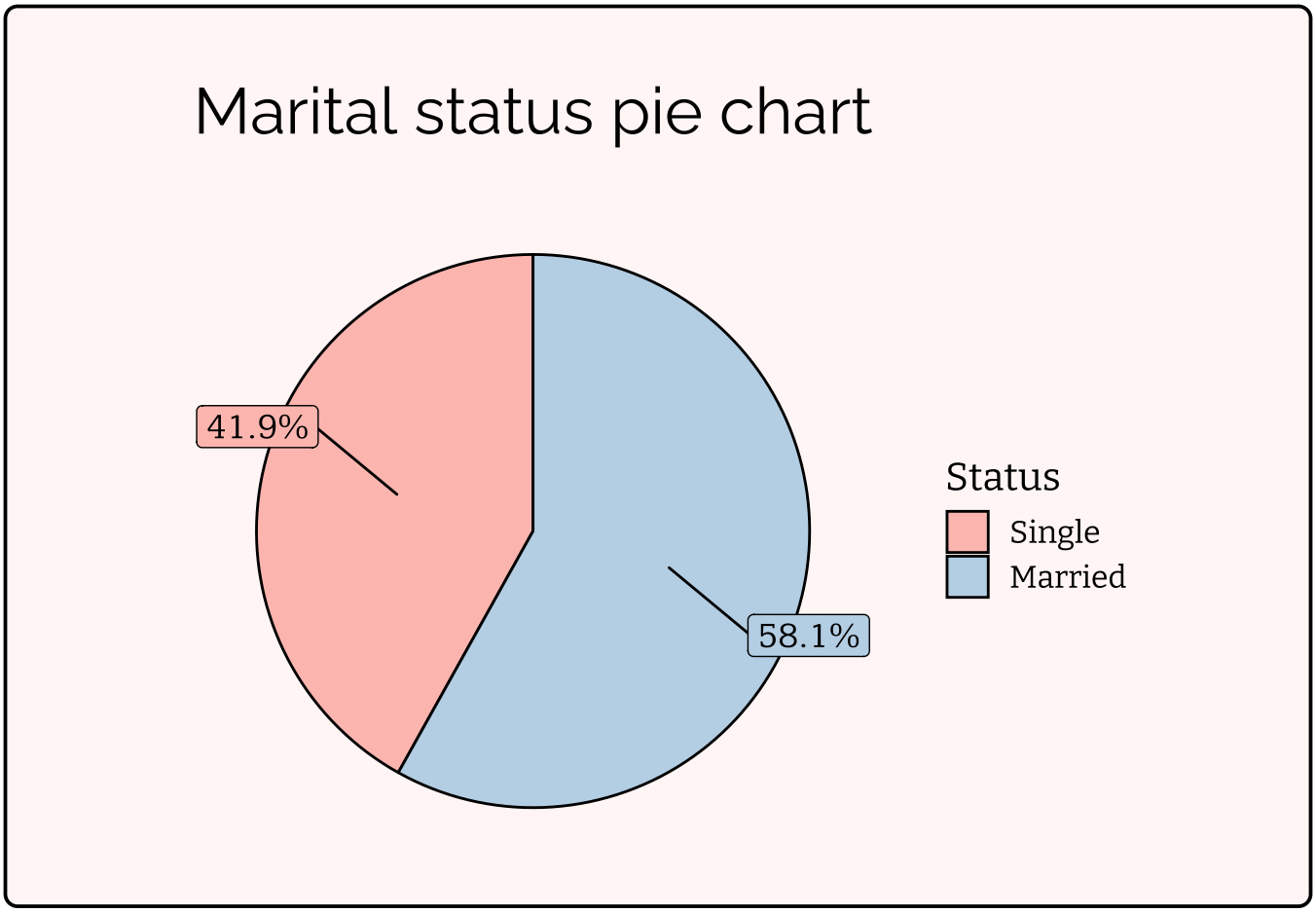
```
married_table <- prop.table(married_table)
married_table

## married
##  Single Married
##   0.419   0.581
```

- Put the processed data into data frame to visualize it.

```
married_table_value <- as.vector(married_table)

df <- data.frame(value = married_table_value * 1e2,
                 group = c("Single", "Married"))
```

- Visualize the data in pie chart to specify the rate between *married* and *single*

```
ggplot(df, aes(x = "", y = value, fill = fct_inorder(group))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Marital status pie chart") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_label_repel(data = df2,
                   family = "Text",
                   aes(y = pos, label = paste0(value, "%")),
                   size = 5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Status")) +
  theme_void() +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.margin = margin(0, 0, 0, 20),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=25, margin=margin(20, 0, 20, 0)),
        text = element_text(family = "Axis", size=15))
```



- **Remark 1:** In the pie chart, it's visually that the percentage of married people is larger than that of single one, with 58.1% and 41.9% respectively.
- Draw box plot to investigate the impact of marital status on the employees' wage

```
ggplot(data, aes(x = married, y = hrswk, fill = married)) +
  geom_boxplot() +
  coord_flip()+
  scale_fill_brewer(palette = "Pastel1", label=c("Single", "Married"),
                    name="Status")+

  scale_x_discrete(labels=c("Single","Married")) +
  labs(title="Plot of weekly working hours by marital status", x="Status",
       y = "Weekly working hours (hours)")+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        legend.position="top",
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=25, margin=margin(20,0,40,0)),

        axis.text = element_text(color="black", family = "Text"),
```

```
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
        text = element_text(family = "Text", size=15, color = "black"))
```

- **Remark 2:**
  - As mentioned above, about 70% employee working 40 hours/week, therefore, the majority of working hour is at 40. It's the reason why it's only the line in box plot for both married and single.
  - Because it's only the line, it's hard to figure out the mean wage of both marital status. We can test it later to make a conclusion about the mean wage of employees.

2.2.2 Attribute *female*

- Count the data and calculate its proportion

```
female_table <- table(female)
female_table

## female
##   Male Female
##    486    514
```

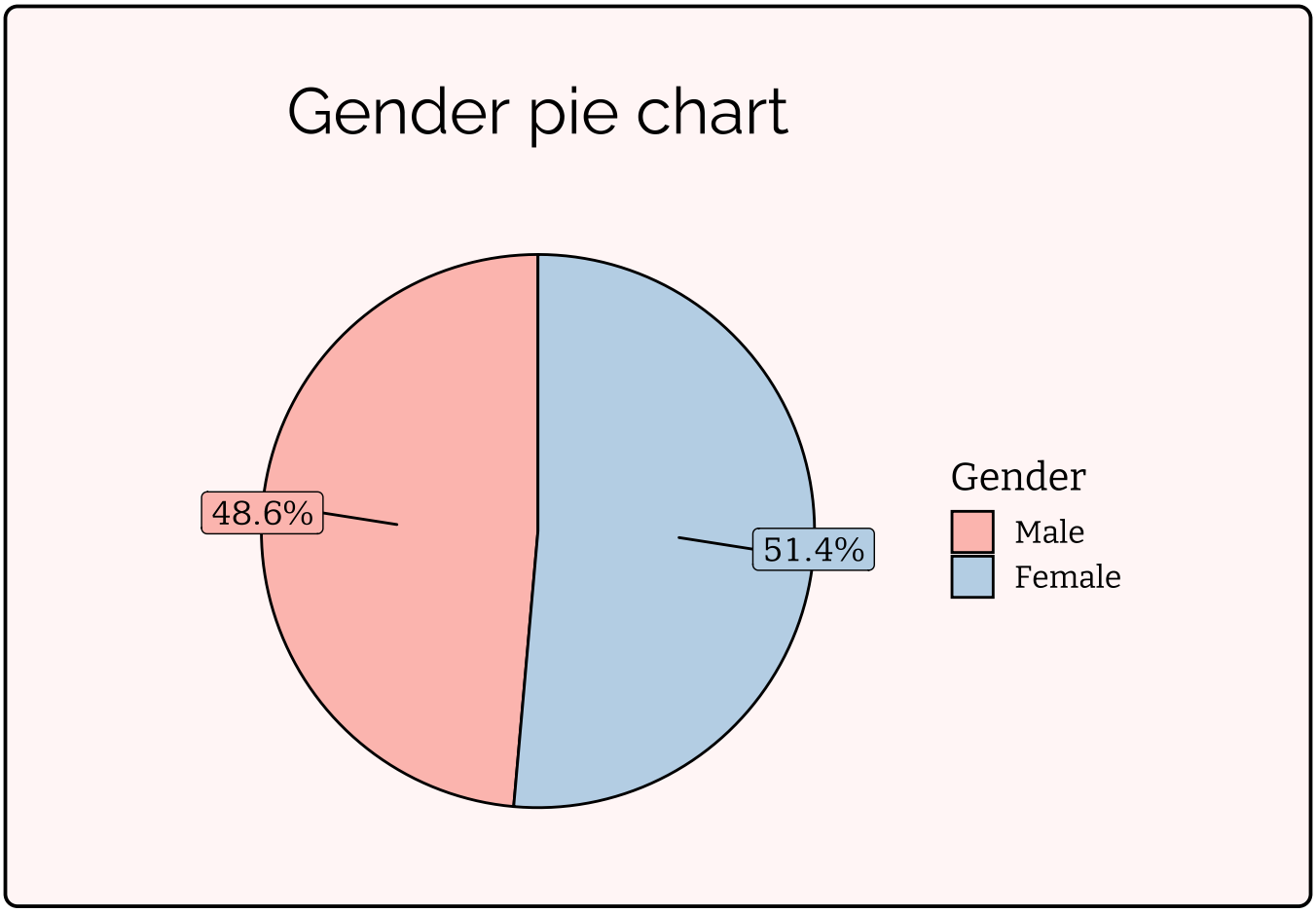
```
female_table_prop <- prop.table(female_table)
female_table_prop

## female
##   Male Female
##  0.486  0.514
```

- Put the processed data into data frame to visualize it.

```
female_table_value_prop <- as.vector(female_table_prop)

df <- data.frame(value = female_table_value_prop * 1e2,
                 group = c("Male", "Female"))
```

- Visualize the data in pie chart to specify the rate between *female* and *male*

```
ggplot(df, aes(x = "", y = value, fill = fct_inorder(group))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Gender pie chart") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_label_repel(data = df2,
                   family = "Text",
                   aes(y = pos, label = paste0(value, "%")),
                   size = 5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Gender")) +
  theme_void() +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.margin = margin(0, 0, 0, 20),
        plot.title = element_text(hjust = 0.5, family = "Title", size=25,
                                  margin=margin(20,0,20,0)),
        text = element_text(family = "Axis", size=15))
```



- **Remark 1:** In the pie chart, though the rate of female is bigger than male, with 51.4% and 48.6% respectively, its rate is quite similar to each other.
- Draw box plot to investigate the impact of gender on the employees' wage

```
get_box_stats <- function(y, upper_limit = max(data$lwage) * 1.15) {

  return(data.frame(
    y = 0.95 * upper_limit,
    label = paste(
      "Count =", length(y), "\n",
      "Mean =", round(summary(y)[4], 2), "\n",
      "Median =", round(summary(y)[3], 2), "\n",
      "First Quartile =", round(summary(y)[2], 2), "\n",
      "Third Quartile =", round(summary(y)[5], 2), "\n"
    )
  ))
}
```

```
ggplot(data, aes(x = female, y = lwage, fill = female)) +
  geom_boxplot() +

  scale_fill_brewer(palette = "Pastel1", labels=c("Male","Female"), name="Gender")+
  stat_summary(fun.data = get_box_stats, geom = "text",
```

```
            vjust = -0.2, hjust = 0.5, family="Text") +
scale_x_discrete(labels=c("Male","Female")) +
ylim(0, 7)+
labs(title="Plot of wage with respect to gender",x="Gender", y = "Lwage ($)")+
theme_classic() +
theme(plot.background = element_blank(),
      panel.background = element_blank(),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      legend.background = element_blank(),
      legend.position="right",
      plot.title = element_text(hjust = 0.5, family = "Title",
                                margin=margin(20,0,40,0), size=25),
      axis.text = element_text(color="black", family = "Text"),
      axis.title = element_text(color="black", family="Axis"),
      axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
      axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
      text = element_text(family = "Text", size=15, color = "black"))
```

- **Remark 2:**
  - In the box plot, the range from first to third quartile of both gender is quite the same, approximately 1, so their box plots are rather similar in size.
  - The mean as well as the median wage of female employees (2.77 and 2.71 respectively) is smaller than that of male (2.95 and 2.96 respectively), but the difference is not too much.
  - We can find out that not only the mean and the median, but also the first and third quartile of male box plot is greater than the female one, which infers that most of male wage is higher than female.
- The relationship between gender rate and marriage rate

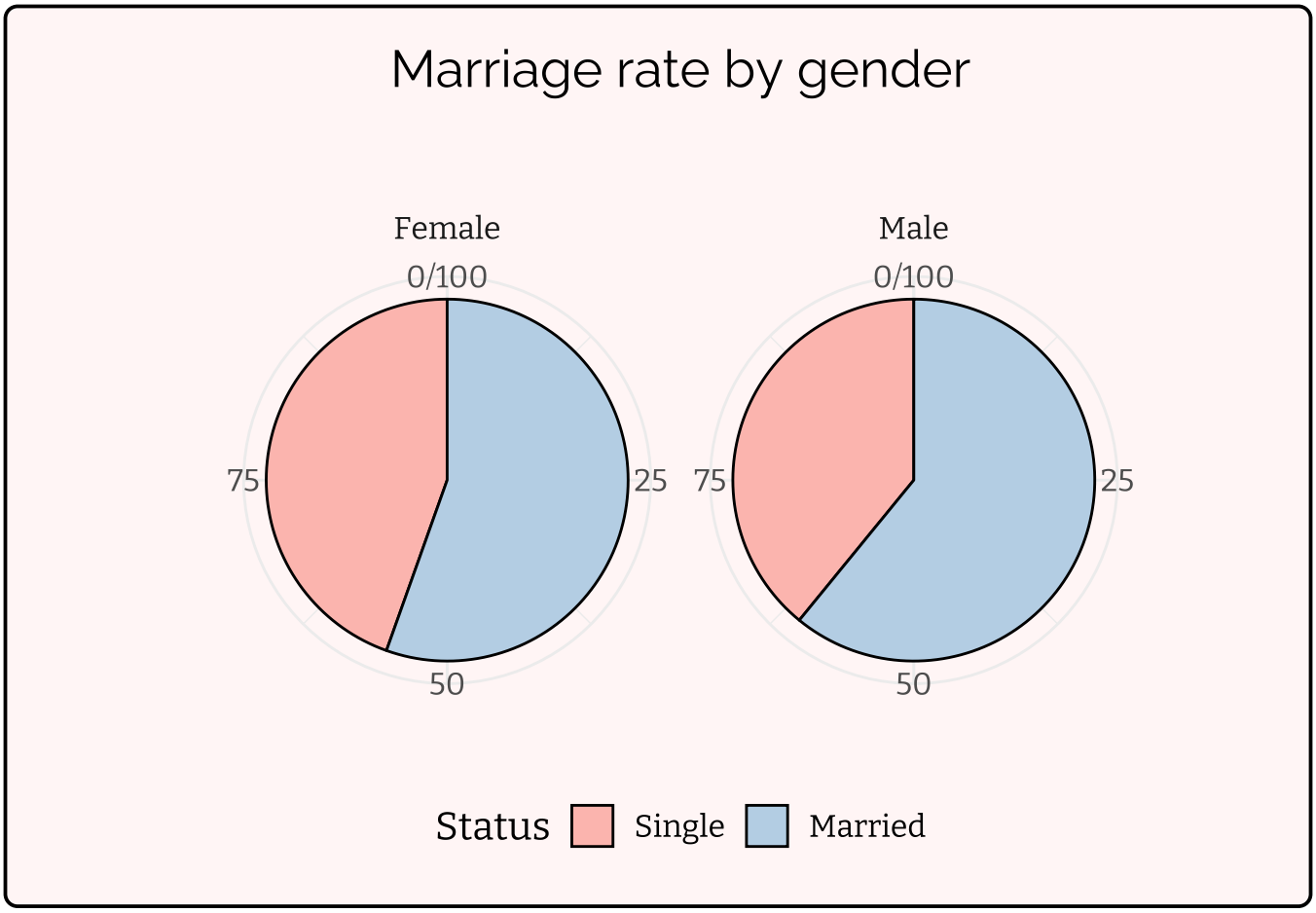
```r
female_married <- data %>%
  group_by(female, married) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 1e2)
female_married
```

```
## # A tibble: 4 x 4
## # Groups:   female [2]
##   female married count percentage
##   <fct>  <fct>   <int>      <dbl>
## 1 Male   Single    190       39.1
## 2 Male   Married   296       60.9
## 3 Female Single    229       44.6
## 4 Female Married   285       55.4
```

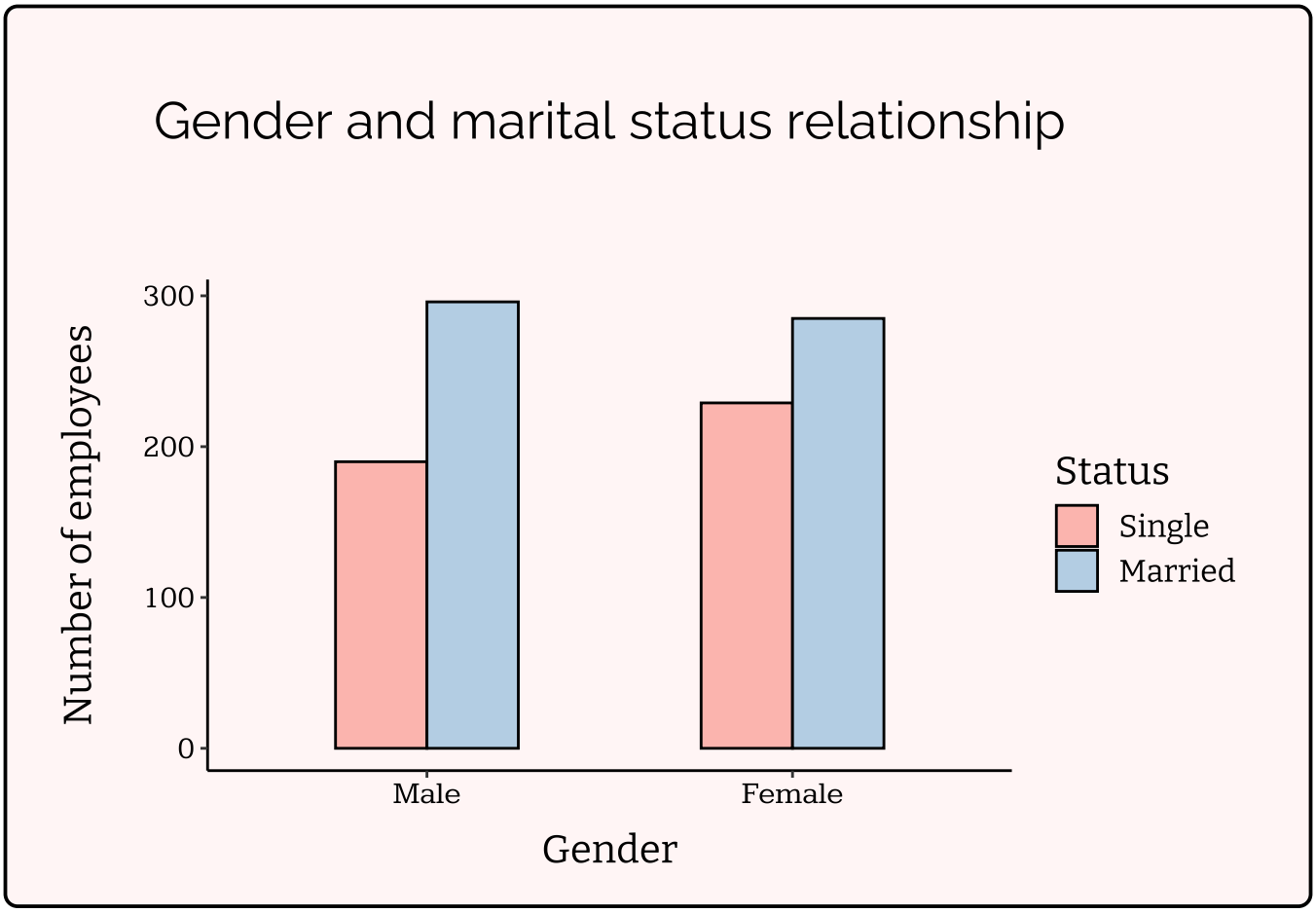
```r
df3 <- data.frame(value = female_married$percentage,
                  group1 = c("Male", "Male", "Female", "Female"),
                  group2 = c("Single", "Married", "Single", "Married"))


df3
```

```
##      value group1  group2
## 1 39.09465   Male  Single
## 2 60.90535   Male Married
## 3 44.55253 Female  Single
## 4 55.44747 Female Married
```

```r
ggplot(df3, aes(x = '', y = value, fill = fct_inorder(group2))) +
  geom_bar(width = 2, stat = 'identity', color=1) +
  coord_polar('y', start = 0) +
  facet_wrap(~group1, ncol = 2, scale = 'fixed') +
  ggtitle('Marriage rate by gender') +
  xlab('') +
  ylab('') +
  scale_fill_brewer(palette = 'Pastel1', name = 'Status',
                    labels = c('Single', 'Married')) +
  theme_minimal() +
  theme(legend.position = "bottom",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=20, margin=margin(0, 0, 40, 0)),
        text = element_text(family = "Axis", size=15))
```




```
ggplot(female_married, aes(x = female, y = count, fill = married)) +
  geom_bar(stat='identity', width=0.5, position="dodge", color="black") +
  scale_fill_brewer(palette = "Pastel1", labels=c("Single", "Married"),
                    name="Status")+
  labs(title="Gender and marital status relationship",
       x = "Gender",
       y = "Number of employees") +
  scale_x_discrete(labels=c("Male","Female")) +
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=20, margin=margin(20,0,50,0)),
        axis.text = element_text(color="black", family = "Text"),
        axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        text = element_text(family = "Axis", size=15))
```



- **Remark 3:**
  - From the above charts, we can realize that more than 50% of employees in both gender got married, which means the married proportion in male is greater the percentage of single status, 60.91% compared to 39.09% and the same as female with 55.45% compared to 44.55%.
  - Another feature that can be inferred from those graphs is the rate of married in male (60.91%) is slightly greater than that of female (55.45%).
  - The distribution of 2 bars in each section looks similar to each other, we will experiment the independence test later to conclude its relationship.

2.2.3 Attribute *metropolitan*

- Count the data and calculate its proportion

```
metro_table <- table(metro)
metro_table
```

```
## metro
## Outskirts     Metro
##       220       780
```

```
metro_table <- prop.table(metro_table)
metro_table
```

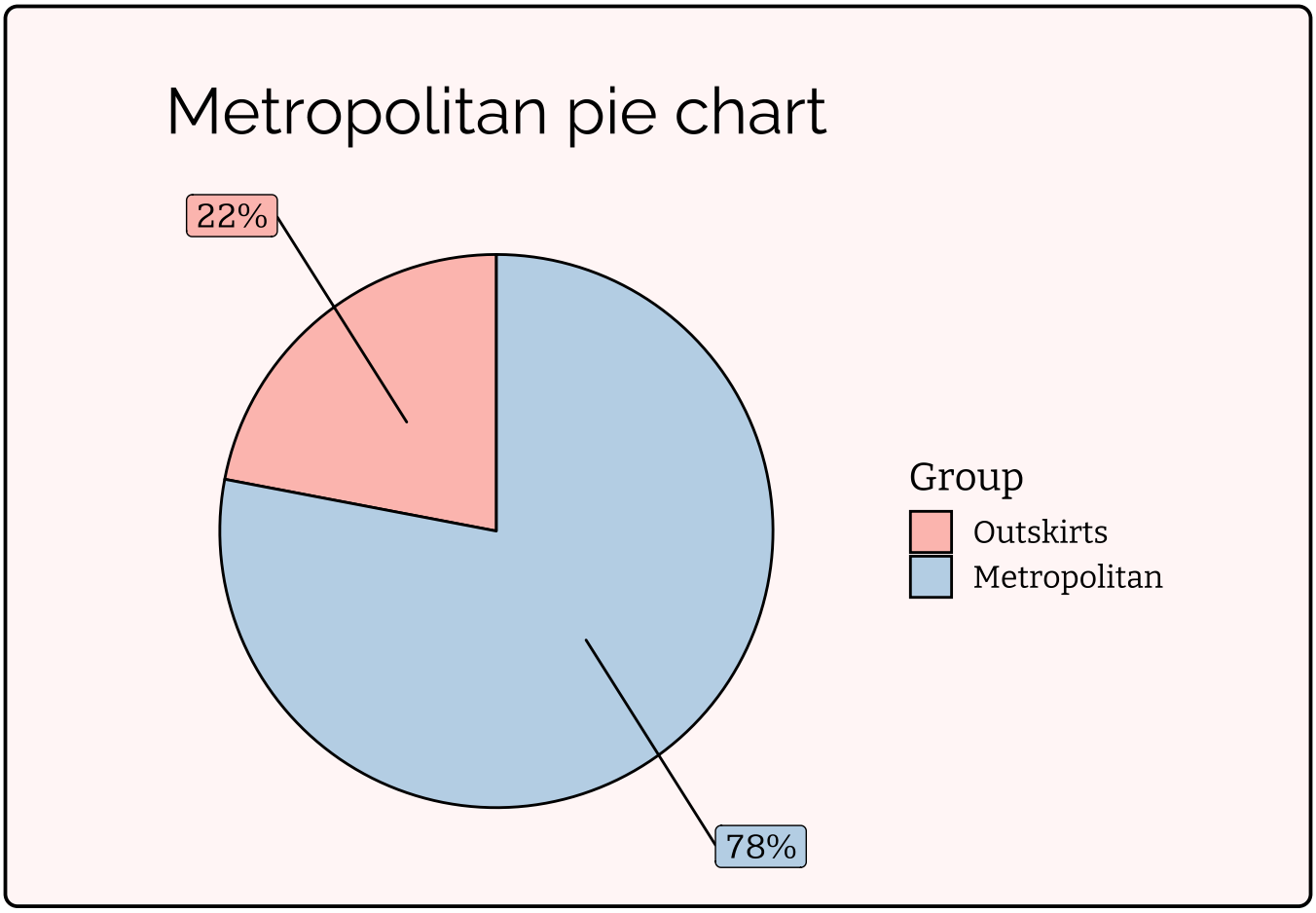
```
## metro
## Outskirts     Metro
##      0.22      0.78
```

- Put the processed data into data frame to visualize it.

```
metro_table_value <- as.vector(metro_table)

df <- data.frame(value = metro_table_value * 1e2,
                 group = c("Outskirts", "Metropolitan"))
```

- Visualize the data in pie chart to specify the rate between employees living in *metropolitan* and *outskirts*

```
ggplot(df, aes(x = "", y = value, fill = fct_inorder(group))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Metropolitan pie chart") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_label_repel(data = df2,
                   family = "Text",
                   aes(y = pos, label = paste0(value, "%")),
                   size = 5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Group")) +
  theme_void() +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.margin = margin(0, 0, 0, 20),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=25, margin=margin(20,0,20,0)),
        text = element_text(family = "Axis", size=15))
```



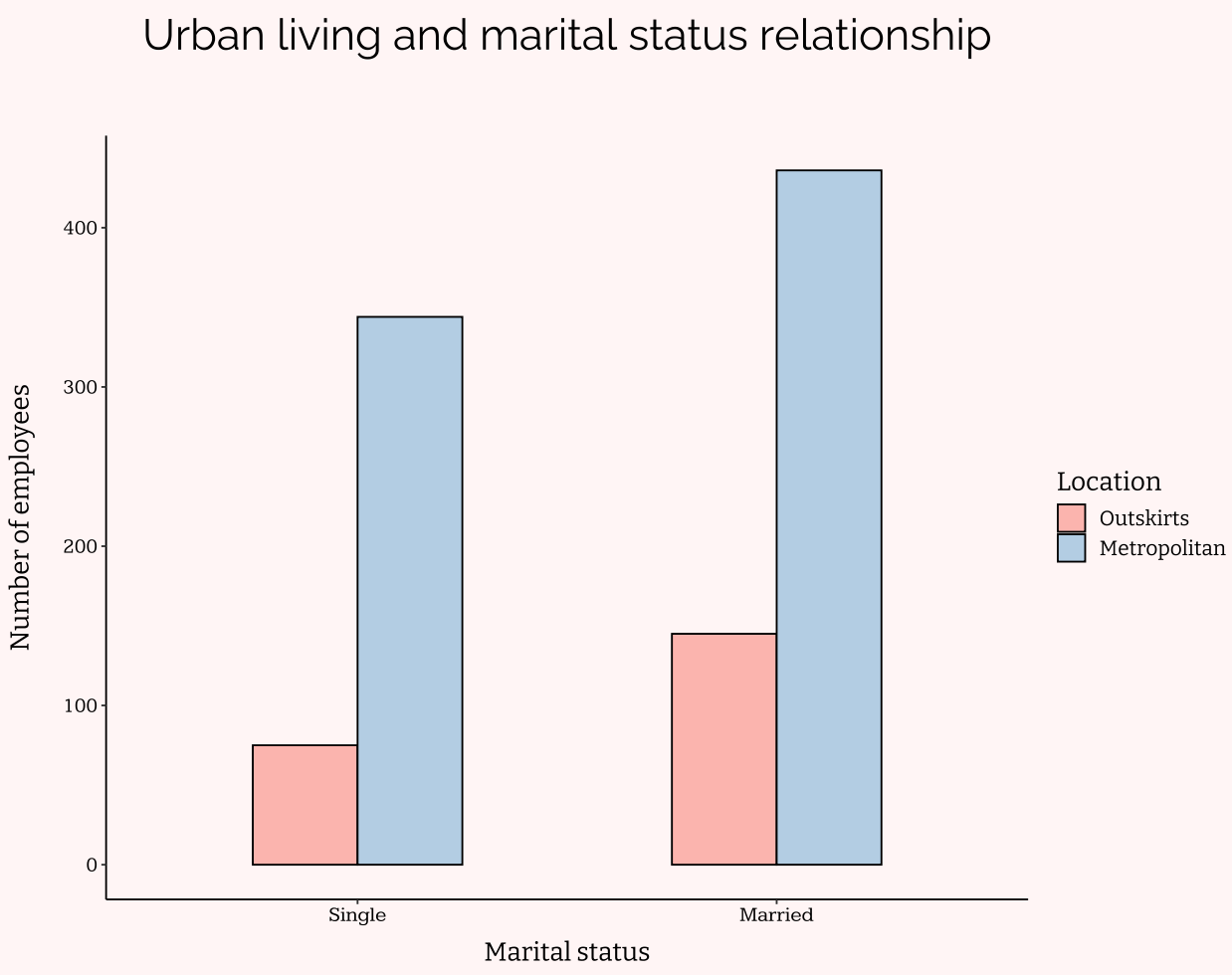
- **Remark 1:** In the pie chart, rate of metropolitan dwellers dominates that of outskirts inhabitants, with 78% and 22%. It implies that most of the employees in the survey are living in big city.
- Draw box plot to investigate whether marriage affects decision to live in civic.

```
married_metro <- data %>%
  group_by(married, metro) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 1e2)
married_metro
```

```
## # A tibble: 4 x 4
## # Groups:   married [2]
##   married metro     count percentage
##   <fct>   <fct>     <int>      <dbl>
## 1 Single  Outskirts    75       17.9
## 2 Single  Metro       344       82.1
## 3 Married Outskirts   145       25.0
## 4 Married Metro       436       75.0
```

```
ggplot(married_metro, aes(x = married, y = count, fill = metro)) +
  geom_bar(stat='identity', width=0.5, position="dodge", color="black") +
  scale_fill_brewer(palette = "Pastel1", labels=c("Outskirts", "Metropolitan"),
                    name="Location")+
  labs(title="Urban living and marital status relationship",
```

```
      x = "Marital status",
      y = "Number of employees") +
 scale_x_discrete(labels=c("Single","Married")) +
 theme_classic() +
 theme(plot.background = element_blank(),
       panel.background = element_blank(),
       panel.grid.major = element_blank(),
       panel.grid.minor = element_blank(),
       legend.background = element_blank(),
       plot.title = element_text(hjust = 0.5, family = "Title",
                                 size=25, margin=margin(20,0,50,0)),
       axis.text = element_text(color="black", family = "Text"),
       axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
       axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
       text = element_text(family = "Axis", size=15))
```



- **Remark 2:**
  - From the above chart, the rate of employees living in metropolitan for any status of marriage all outweigh the rate for living in suburbs.
  - Because the number of married worker is greater than the quantity of single employee, therefore, both 2 columns in section married are all higher than columns in single one.

  - The distribution of 2 bars in each section looks similar to each other, we will experiment the independence test later to conclude its relationship.

- Examine the relationship of *asian* and *metro*

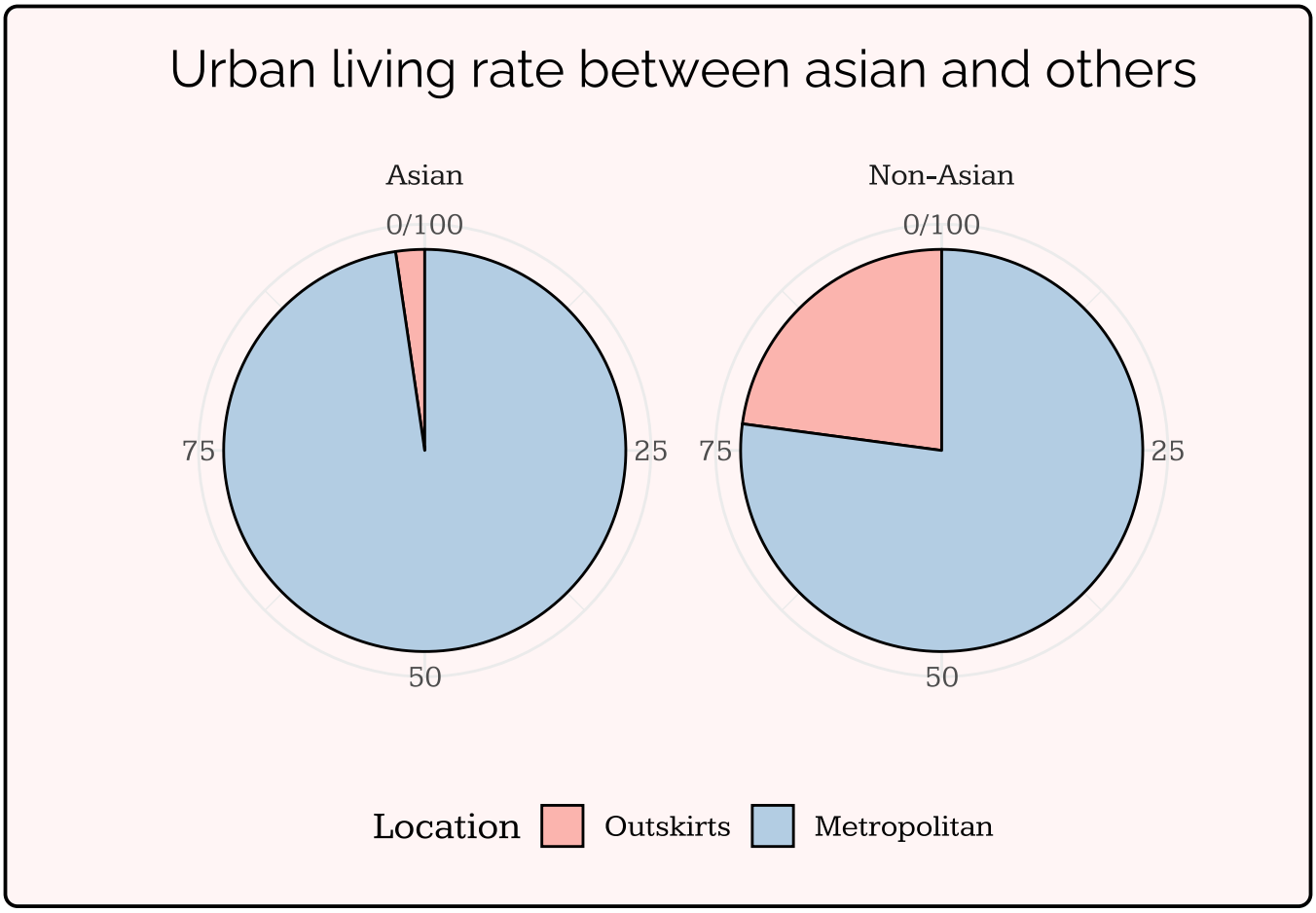
```
asian_metro <- data %>%
  group_by(asian, metro) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 1e2)

asian_metro <- asian_metro[1:4,]
asian_metro
```

```
## # A tibble: 4 x 4
## # Groups:   asian [2]
##   asian metro     count percentage
##   <fct> <fct>     <int>      <dbl>
## 1 0     Outskirts   219      22.9
## 2 0     Metro       738      77.1
## 3 1     Outskirts     1       2.33
## 4 1     Metro        42      97.7
```

```
df4 <- data.frame(value = asian_metro$percentage,
                  group1 = c("Non-Asian", "Non-Asian", "Asian", "Asian"),
                  group2 = c("Outskirts", "Metropolitan",  "Outskirts",
                             "Metropolitan"))


df4
```

```
##       value    group1       group2
## 1 22.884013 Non-Asian    Outskirts
## 2 77.115987 Non-Asian Metropolitan
## 3  2.325581     Asian    Outskirts
## 4 97.674419     Asian Metropolitan
```

```
ggplot(df4, aes(x = '', y = value, fill = fct_inorder(group2))) +
  geom_bar(width = 2, stat = 'identity', color=1) +
  coord_polar('y', start = 0) +
  facet_wrap(~group1, ncol = 2, scale = 'fixed') +
  ggtitle('Urban living rate between asian and others') +
  xlab('') +
  ylab('') +
  scale_fill_brewer(palette = 'Pastel1', name = 'Location',
                    labels = c('Outskirts', 'Metropolitan')) +
  theme_minimal() +
  theme(axis.line = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=20, margin=margin(0, 0, 20, 0)),
        text = element_text(family = "Text", size=15))
```



- **Remark 3:**
  - The pie graph shows that the proportion of living in urban no matter their races dominates the rate of choosing to live in rural areas.
  - It seems like most of asian people decides to live in city center, the rate is nearly 100%, and it's greater than the rate of any other races.

2.2.4 Attribute *midwest*, *south*, and *west*

- Because three attribute *midwest*, *south*, and *west* all consider the employees' area of living, therefore, we combine them into one variable, called *area*, in order to compare them easier.

```
df1 <- data.frame(m = midwest, s = south, w=west, wage = wage)

for(i in 1:nrow(df1)){
  df1$area[i] <- NA
  if(df1$m[i]== 1){
    df1$area[i] <- "m"
  } else if (df1$s[i] == 1) {
    df1$area[i] <- "s"
  } else if (df1$w[i] == 1){
    df1$area[i] <- "w"
  } else{
    df1$area[i] <- "z"
  }
```

```
}

head(df1)

##   m s w  wage area
## 1 0 1 0 18.70    s
## 2 1 0 0 11.50    m
## 3 0 0 1 15.04    w
## 4 0 1 0 25.95    s
## 5 0 0 0 24.03    z
## 6 0 0 0 20.00    z
```

- Count the data and calculate its proportion

```
area_table <- table(df1$area)
area_table

##
##   m   s   w   z
## 240 296 240 224
```

```
area_table <- prop.table(area_table)
area_table

##
##     m     s     w     z
## 0.240 0.296 0.240 0.224
```

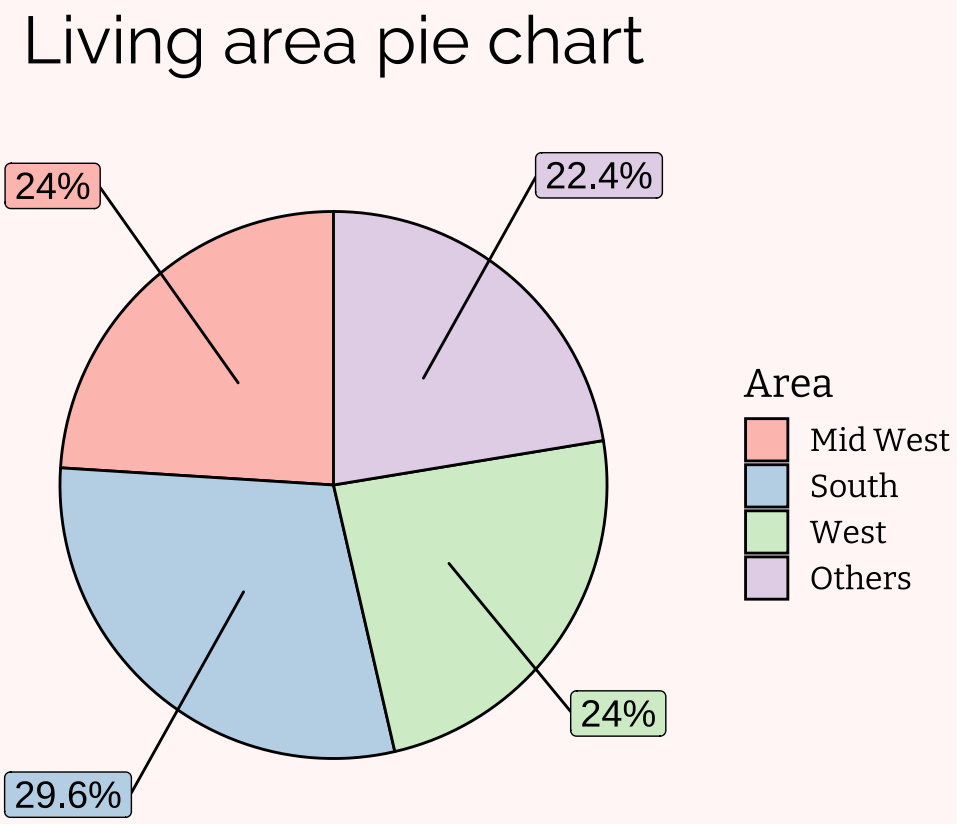
- Put the processed data into data frame to visualize it.

```
area_table_value <- as.vector(area_table)

df <- data.frame(value = area_table_value * 1e2,
                 group = c("Mid West", "South", "West", "Others"))
```

- Visualize the data in pie chart to specify the rate between employees living in *midwest*, *south*, *west* and *others*

```
ggplot(df, aes(x = "", y = value, fill = fct_inorder(group))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Living area pie chart") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(value, "%")),
                   size = 5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Area")) +
  theme_void() +
  theme(axis.line = element_blank(),
```

```
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.margin = margin(0, 0, 0, 20),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=25, margin=margin(20,0,30,0)),
        text = element_text(family = "Axis", size=15))
```

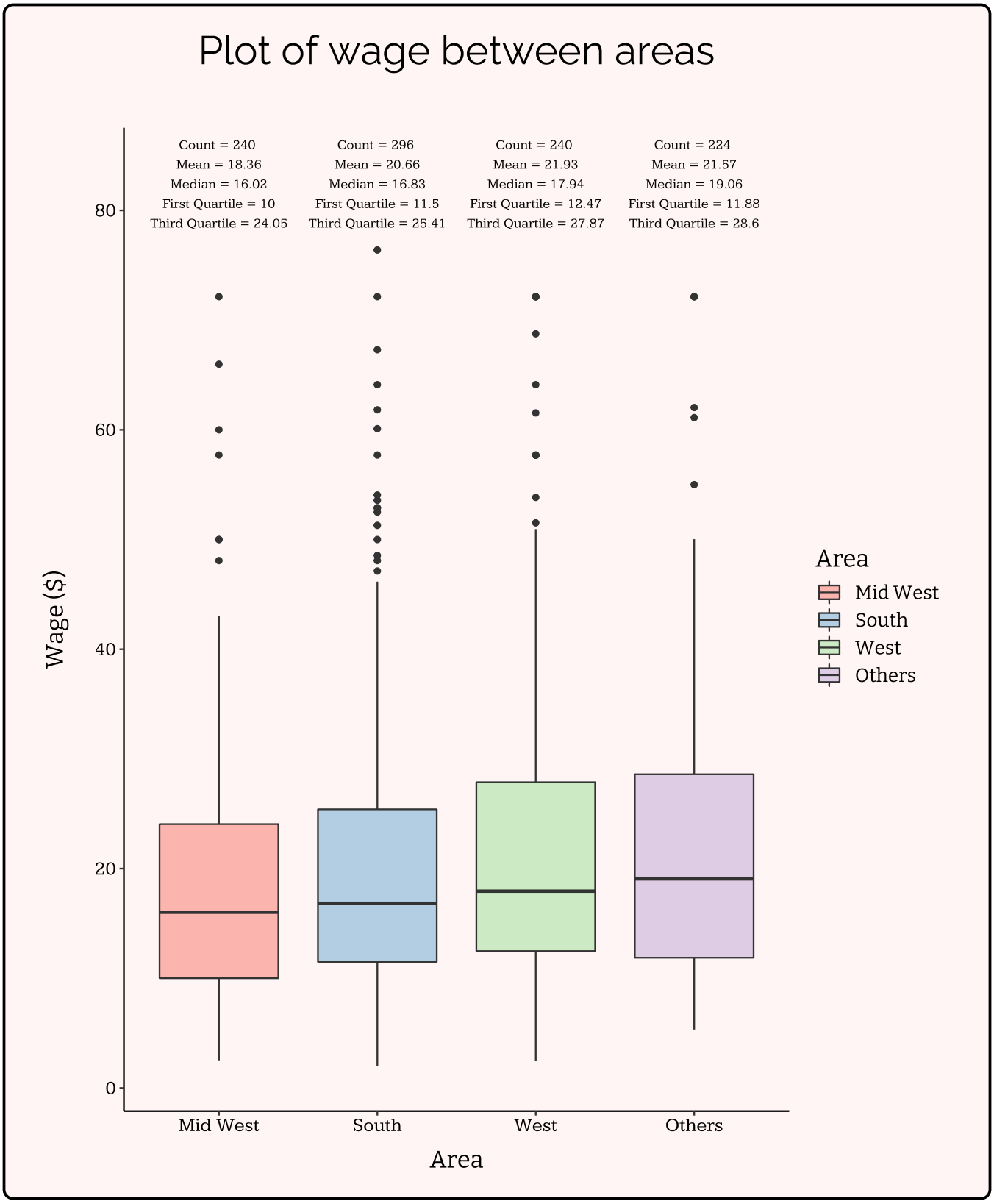


- **Remark 1:**
  - In the pie chart, the four area percentage is almost the same, which is nearly a quarter.
  - The area that the employees live in the most is South, with 29.6%. 24% of workers live in Mid West and West area. The remaining employees, 22.4% live in other areas.
- Draw box plot to investigate how the employees' living area affects their wage

```
get_box_stats <- function(y, upper_limit = max(data$wage) * 1.15) {

  return(data.frame(
    y = 0.95 * upper_limit,
    label = paste(
      "Count =", length(y), "\n",
      "Mean =", round(summary(y)[4], 2), "\n",
      "Median =", round(summary(y)[3], 2), "\n",
      "First Quartile =", round(summary(y)[2], 2), "\n",
      "Third Quartile =", round(summary(y)[5], 2), "\n"
    )
```

```
  ))
}
```

```
ggplot(df1, aes(x = area, y = wage, fill = area)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1",
                    labels=c("Mid West", "South", "West", "Others"), name="Area")+
  stat_summary(fun.data = get_box_stats, geom = "text",
               vjust = 0.7, hjust = 0.5, family="Text", size=3) +
  scale_x_discrete(labels=c("Mid West", "South", "West", "Others")) +
  labs(title="Plot of wage between areas",x="Area", y = "Wage ($)")+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        legend.position="right",
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  margin=margin(0,0,40,0), size=25),
        axis.text = element_text(color="black", family = "Text"),
        axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        text = element_text(family = "Axis", size=15, color = "black"))
```



- **Remark 2:**
  - The four box plots are quite similar in size, as well as its statistical parameter.
  - The largest boxplot is *Others* and smallest is *Mid West*. However, the mean, median and first quartile of four box plots are approximately the same, especially the mean of West and Others. It seems like the

area of living does not affect the employees' wage heavily.

2.2.5 Attribute *black* and *asian*

- Because *black* and *asian* attribute all refers to the employees' race, therefore, we combine them into one variable, called *race*, so that we can compare them easier.

```
df1 <- data.frame(b = black, a <- asian, wage = wage, female)

for(i in 1:nrow(df1)){
  df1$race[i] <- NA
  if(df1$b[i]== 1){
    df1$race[i] <- "b"
  } else if (df1$a[i] == 1) {
    df1$race[i] <- "a"
  }else{
    df1$race[i] <- "o"
  }
}

head(df1)

##   b a....asian  wage female race
## 1 0          0 18.70 Female    o
## 2 0          0 11.50   Male    o
## 3 1          0 15.04   Male    b
## 4 1          0 25.95 Female    b
## 5 0          0 24.03   Male    o
## 6 0          0 20.00   Male    o
```

- Count the data and calculate its proportion

```
race_table <- table(df1$race)
race_table

##
##   a   b   o
##  43 112 845
```

```
race_table <- prop.table(race_table)
race_table

##
##     a     b     o
## 0.043 0.112 0.845
```

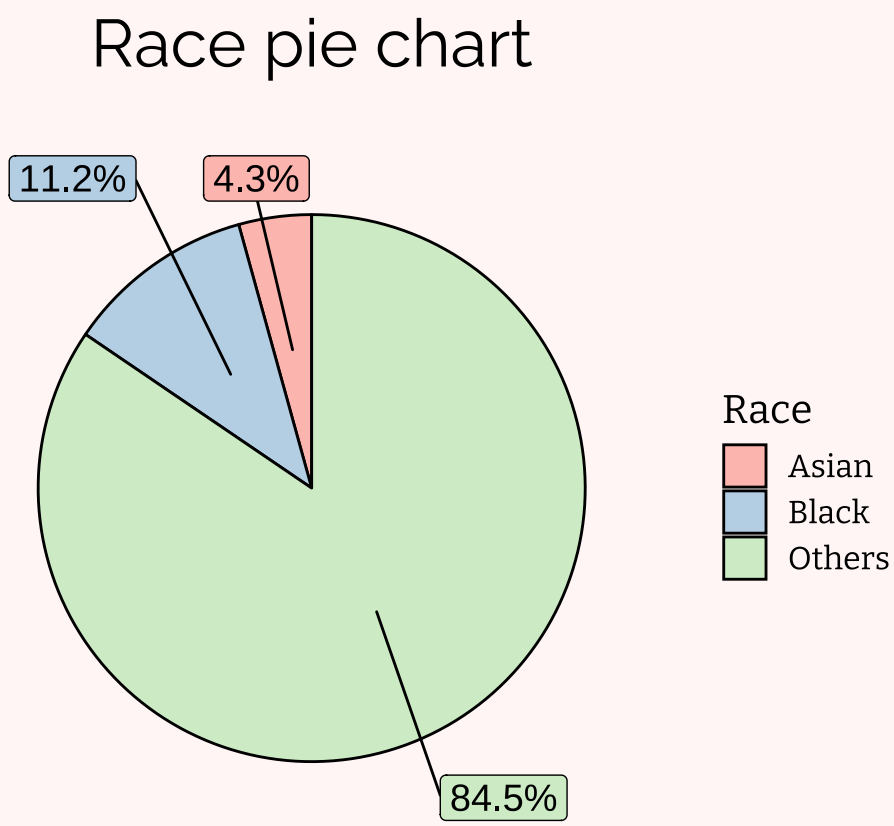
- Put the processed data into data frame to visualize it.

```
race_table_value <- as.vector(race_table)
```

```r
df <- data.frame(value = race_table_value * 1e2,
                 group = c("Asian", "Black", "Others"))
```

- Visualize the data in pie chart to specify the rate between *black*, *asian* and *others* race employee

```r
ggplot(df, aes(x = "", y = value, fill = fct_inorder(group))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  labs(x = NULL, y = NULL, fill = NULL,
       title = "Race pie chart") +
  scale_fill_brewer(palette = "Pastel1")+
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(value, "%")),
                   size = 5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Race")) +
  theme_void() +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        legend.margin = margin(0, 0, 0, 20),
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  size=25, margin=margin(20,0,30,0)),
        text = element_text(family = "Axis", size=15))
```



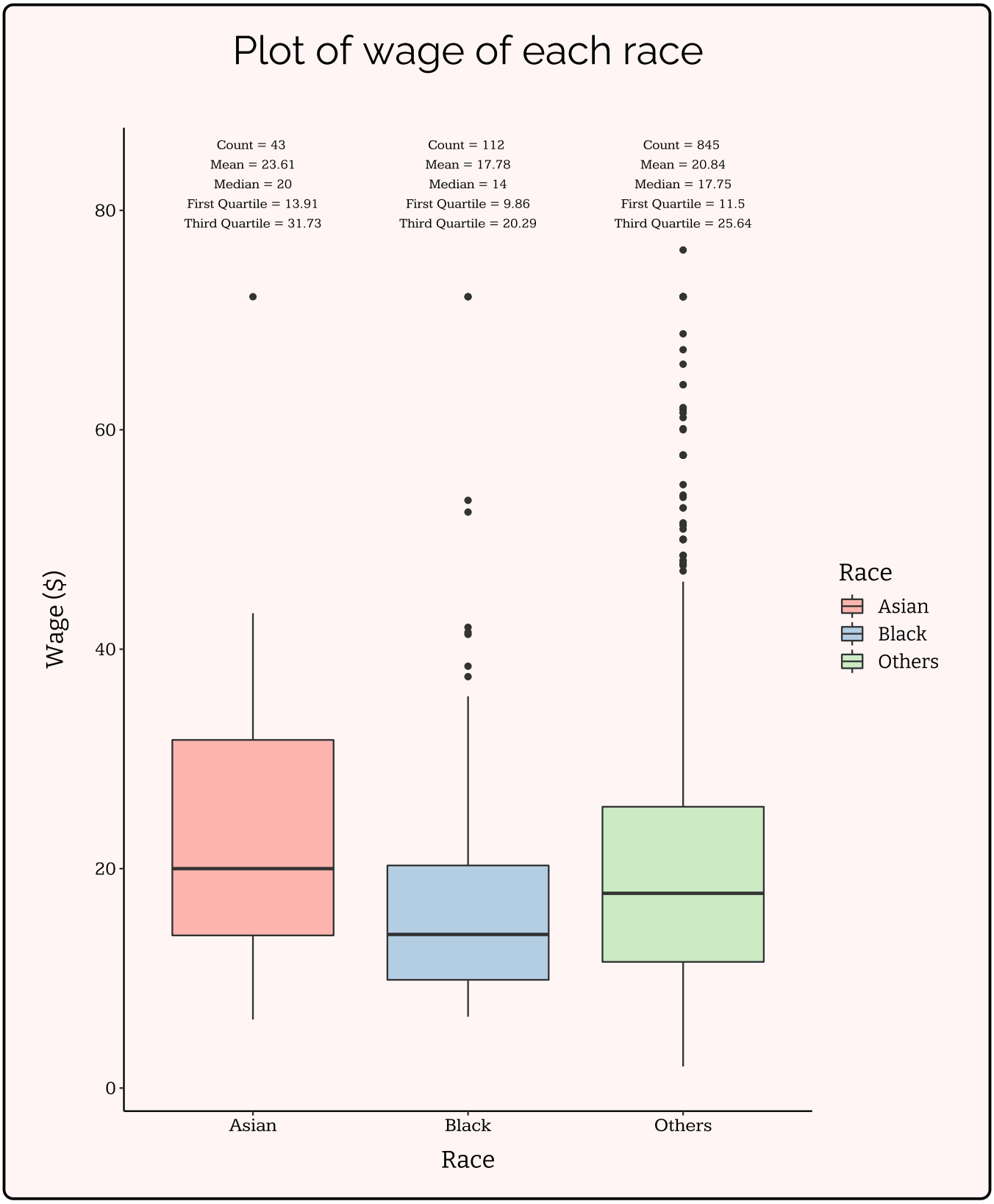
- **Remark 1:**

- In the pie chart, the *Others* race percentage surpass the remaining race which is *Asian* and *Black*.
- Most of the employees in the dataset is not *Asian* and *Black*, mostly White people, which occupies 84.5%. 11.2% workers are black people, the second-largest race. The last and the least one, Asian people, only takes 4.3% in total.

- Draw box plot to investigate how the employees' race affects their wage

```
get_box_stats <- function(y, upper_limit = max(data$wage) * 1.15) {

  return(data.frame(
    y = 0.95 * upper_limit,
    label = paste(
      "Count =", length(y), "\n",
      "Mean =", round(summary(y)[4], 2), "\n",
      "Median =", round(summary(y)[3], 2), "\n",
      "First Quartile =", round(summary(y)[2], 2), "\n",
      "Third Quartile =", round(summary(y)[5], 2), "\n"
    )
  ))
}
```

```
ggplot(df1, aes(x = race, y = wage, fill = race)) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Pastel1",
                    labels=c("Asian", "Black", "Others"), name="Race")+

  stat_summary(fun.data = get_box_stats, geom = "text",
               vjust = 0.7, hjust = 0.5, family="Text", size=3) +
  scale_x_discrete(labels=c("Asian", "Black", "Others")) +
  labs(title="Plot of wage of each race", x="Race", y = "Wage ($)")+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        legend.position="right",
        plot.title = element_text(hjust = 0.5, family = "Title",
                                  margin=margin(0,0,40,0), size=25),
        axis.text = element_text(color="black", family = "Text"),
        axis.title.y = element_text(margin=margin(0, 20, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        text = element_text(family = "Axis", size=15, color = "black"))
```



- **Remark 2:**
  - The 4 box plot are uneven in size. Asian people has the least number but its box is the biggest. It implies that most of asian wage range is mainly in the range from 13.91$ to 31.73$. The Black box plot is rather short compares to others.

- We can see that, there are a lot of outliers in Others above the box, which can be inferred that the wage of Whites varies a lot, and much higher than the two remaining races.
- The median of the Asian box plot is the same as the third quartile of Black box plot, which means 50% of the Asian wage is similar to 75% of Black one.

3 Inferential Statistics

Based on the descriptive statistics, we will pose some statistical problems about topical issue to specify the relationship between the wage and other attributes

3.1 Hypothesis testing for mean

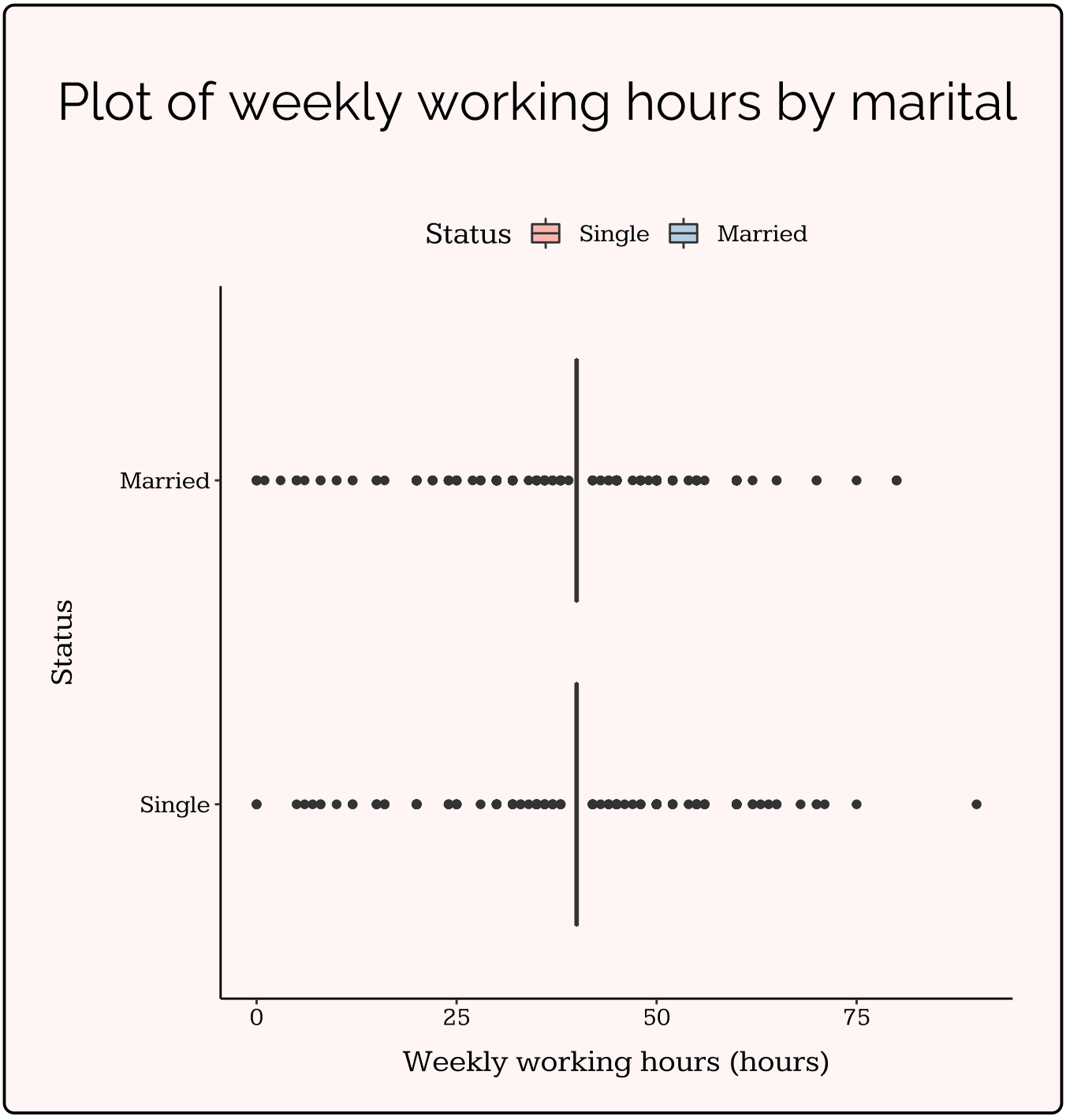
3.1.1 Attribute *wage* with respect to *female*

- **Problem:** Gender discrimination in compensation and promotion has long been a concern of society. From the boxplot above, we see that, the mean of male's wage is a bit bigger than female. Therefore, we will use `t.test` to investigate whether the mean wage of male is higher than that of female to evaluate the gender equality among 1000 employees.
  - Null hypothesis: The mean wage of male and female is the same
  - Alternative hypothesis: The mean wage of male is greater than female

```
t.test(lwage ~ female, data = data, alternative = 'greater')

##
## Welch Two Sample t-test
##
## data: lwage by female
## t = 4.8329, df = 997.94, p-value = 7.785e-07
## alternative hypothesis: true difference in means between group Male and group Female
is greater than 0
## 95 percent confidence interval:
## 0.1155697 Inf
## sample estimates:
## mean in group Male mean in group Female
## 2.947082 2.771801
```

- **Conclusion:** Because the $p-value$ is $7.785e-07$, which is much less than *alpha* (0.05 as default), so we reject the null hypothesis. Therefore, there is sufficient evidence to support the alternative hypothesis that the mean wage of male is greater than that of female.

3.1.2 Attribute *hrswk* with respect to *married*



- **Problem:** There is a popular thinking that married people will spend more time with their family, which means they will work fewer hours than the single one. In the chart above, the box plots for both married and single form a line, it's hard to state anything about its mean. Therefore, we will implement the `t.test` to inspect whether the mean working hours of single employee is more than married one.
  - Null hypothesis: The mean working hours of both married and single worker are equal
  - Alternative hypothesis: The mean working hours of single is greater than married

```
t.test(hrswk ~ married, data = data, alternative = 'greater')

##
```

```
## Welch Two Sample t-test
##
## data: hrswk by married
## t = 0.63485, df = 877.66, p-value = 0.2628
## alternative hypothesis: true difference in means between group Single and group
Married is greater than 0
## 95 percent confidence interval:
## -0.6750187 Inf
## sample estimates:
## mean in group Single mean in group Married
## 40.19809 39.77453
```

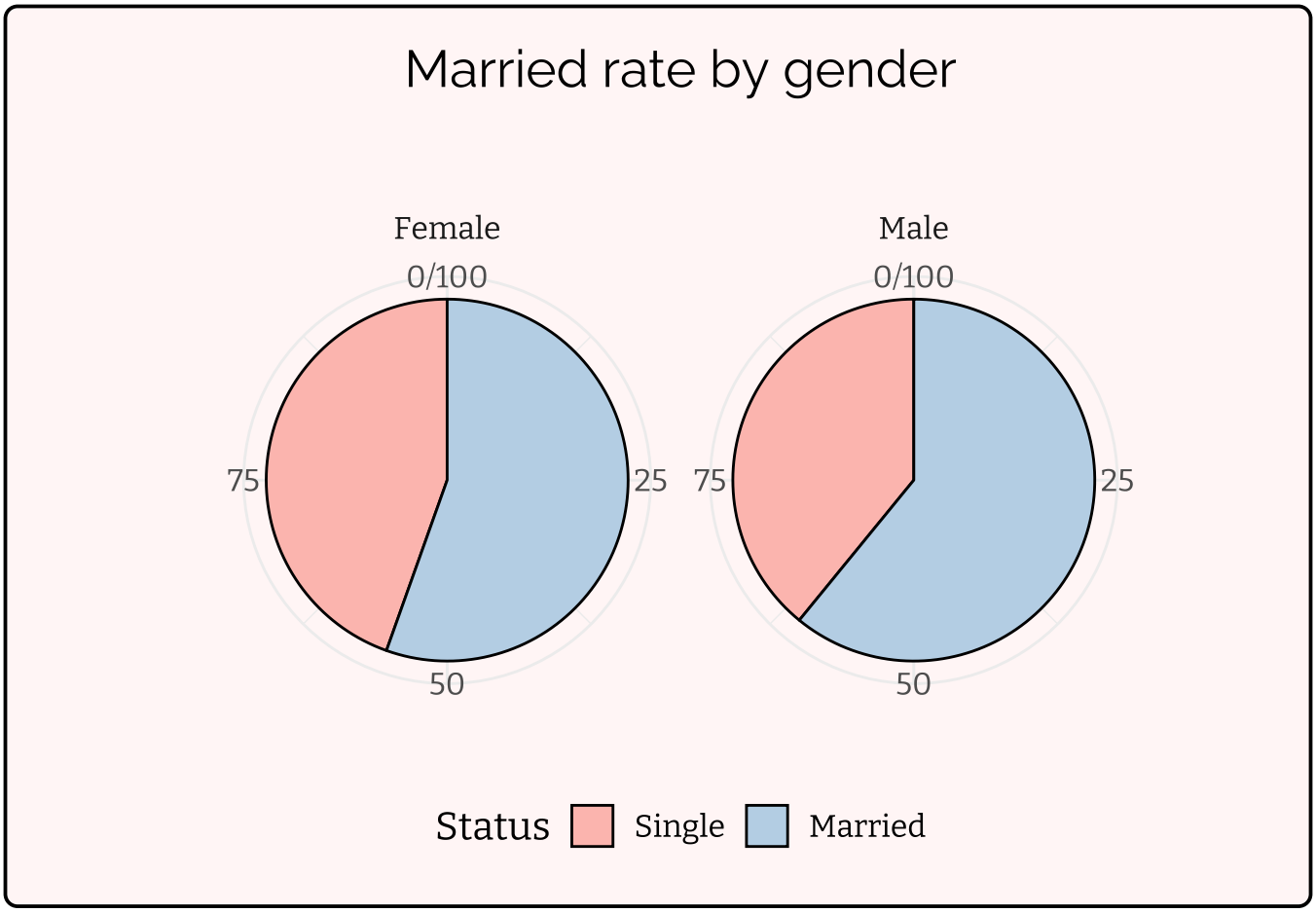
- **Conclusion:** Because $p-value = 0.2628$ which is larger than $\alpha = 0.05$, we accept null hypothesis. Thus, there is not enough evidence to support the hypothesis that single employees work more hours than the married one.

3.2 Hypothesis testing for proportion

3.2.1 Attribute *married* with respect to *female*

```
female_married
```

```
## # A tibble: 4 x 4
## # Groups:   female [2]
##   female married count percentage
##   <fct>  <fct>   <int>      <dbl>
## 1 Male   Single    190       39.1
## 2 Male   Married   296       60.9
## 3 Female Single    229       44.6
## 4 Female Married   285       55.4
```



- **Problem:** As we see in the above graph, the marriage proportion of female is a bit less than male. So we will carry out the test to find out.
  - Null hypothesis: The marriage rate of both gender is equal
  - Alternative Hypothesis: The marriage rate of male is greater than male

```
gender_married <- table(female, married)
gender_married <- gender_married[,c("Married", "Single")]
gender_married
```

```
##         married
## female   Married Single
##   Male       296    190
##   Female     285    229
```

```
prop.test(gender_married, correct = FALSE, alternative = 'greater')
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  gender_married
## X-squared = 3.0567, df = 1, p-value = 0.0402
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.003335837 1.000000000
```

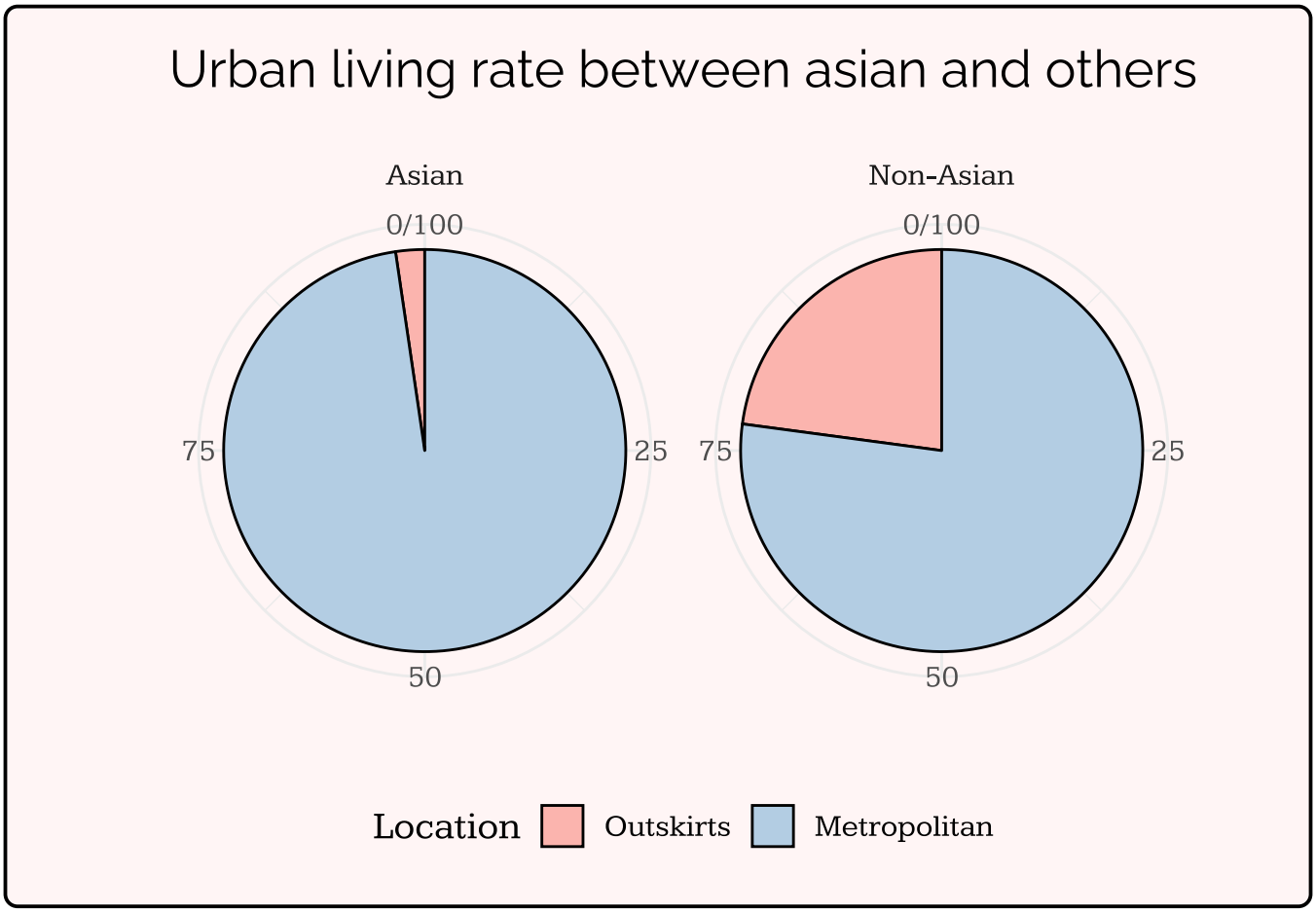
```
## sample estimates:
##    prop 1    prop 2
## 0.6090535 0.5544747
```

- **Conclusion:** Because $p - value = 0.0402$ which is less than $\alpha = 0.05$, we reject null hypothesis. With `prop 1` is the marriage rate of male and `prop 2` is marriage rate of female, there is enough evidence to support the hypothesis that male's rate of marriage is bigger than female.

3.2.2 Attribute *metropolitan* with respect to *asian*

```
asian_metro
```

```
## # A tibble: 4 x 4
## # Groups:   asian [2]
##   asian metro     count percentage
##   <fct> <fct>     <int>      <dbl>
## 1 0     Outskirts   219      22.9
## 2 0     Metro       738      77.1
## 3 1     Outskirts     1       2.33
## 4 1     Metro        42      97.7
```



- **Problem:** As we see in the above graph, the proportion of Asian dwelling in city center is greater than other races. So we will carry out the test to check that.

- Null hypothesis: The urban living rate of both all races is similar to each other
- Alternative Hypothesis: The urban living rate of Asian is greater than other races.

```
asian_metro_table <- table(asian, metro)
rownames(asian_metro_table) = c('Non-asian', 'Asian')
asian_metro_table <- asian_metro_table[c('Asian', 'Non-asian'),
                                       c("Metro", "Outskirts")]

asian_metro_table
```

```
##            metro
## asian       Metro Outskirts
##   Asian        42         1
##   Non-asian   738       219
```

- Using `prop.testcheck`

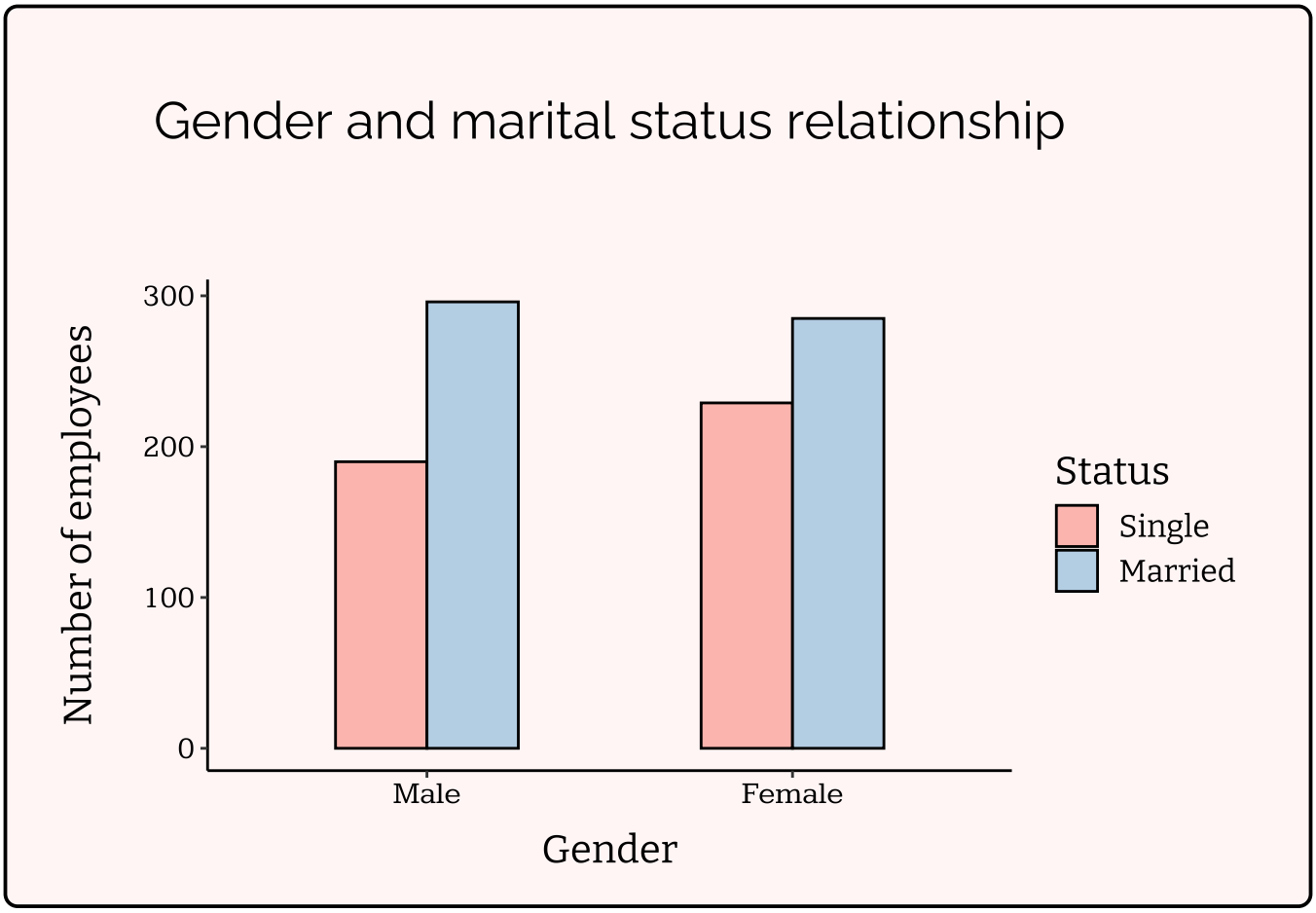
```
prop.test(asian_metro_table, correct=FALSE, alternative = 'greater')
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  asian_metro_table
## X-squared = 10.135, df = 1, p-value = 0.0007272
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1616739 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.9767442 0.7711599
```

- **Conclusion:** Because $p - value = 0.0007272$ which is less than $\alpha = 0.05$, we reject null hypothesis. Thus, there is enough evidence to support the hypothesis that the proportion of Asian living in urban is larger than Non-Asian.

3.3 Test of Independence

3.3.1 Attribute *gender* and *married*

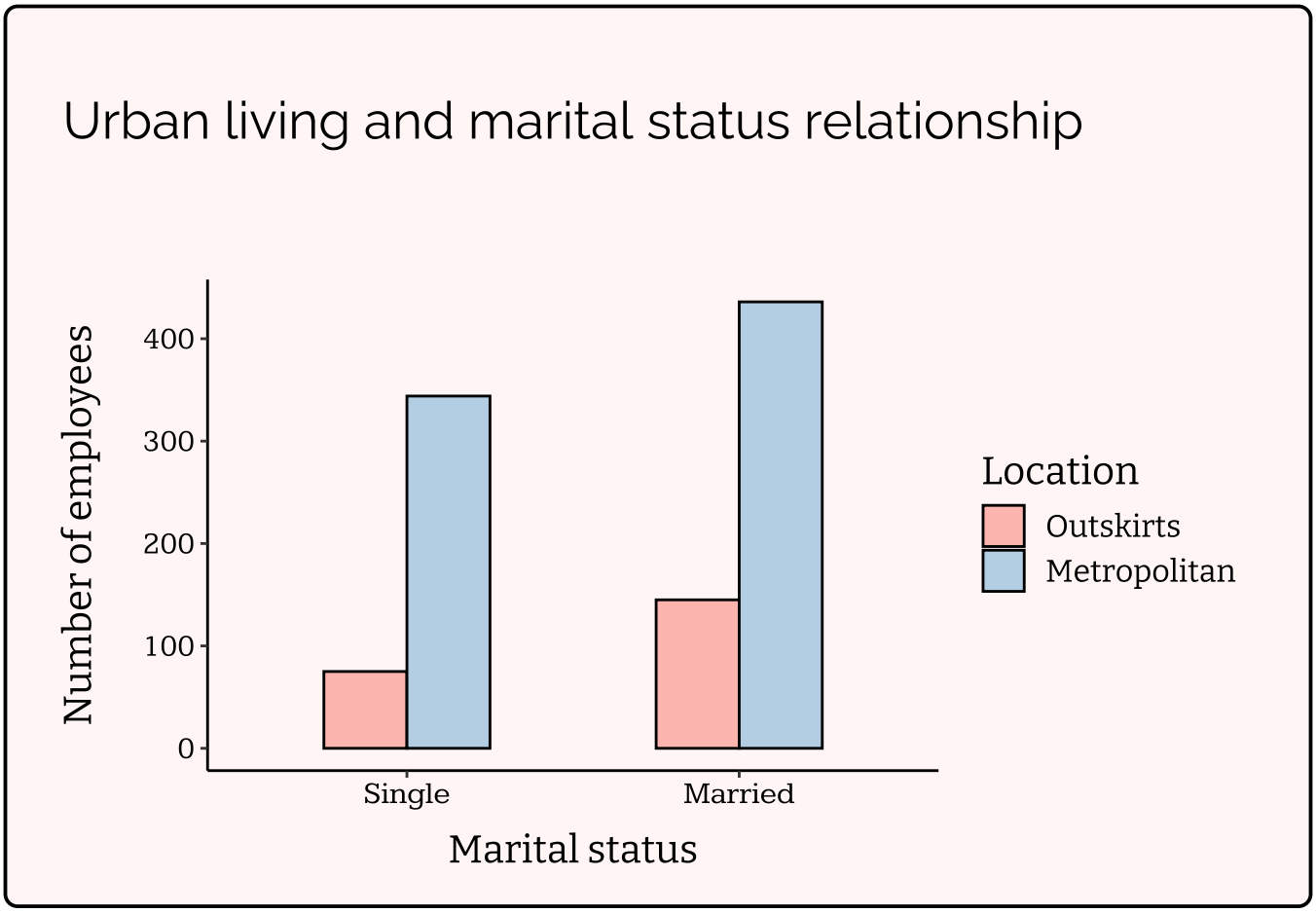


- **Problem:** From the graph above, we can see that the marriage rate of both gender is greater than the single rate. Therefore, it's not clear whether the marriage relates to gender or not. We will implement the independence test (`chisq.test`) to examine its relationship.

```
chisq.test(gender_married)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  gender_married
## X-squared = 2.8366, df = 1, p-value = 0.09214
```

- **Conclusion:** With $p-value = 0.09214$ which is greater than $\alpha = 0.05$, we accept null hypothesis. Thus, there is not enough evidence to support the hypothesis that the marriage relates to gender.

3.3.2 Attribute *married* and *metro*



- **Problem:** From the graph above, we can see that the rate of living in metro in any marital status is greater than the rate of living in the suburbs. Therefore, it's not clear whether the choices of living in urban relates to marital status or not. We will implement the independence test (**`chisq.test`**) to examine its relationship.

```
metro_married <- table(married, metro)

rownames(metro_married) = c('Single', 'Married')
colnames(metro_married) = c('Outskirts', 'Metro')

metro_married
```

```
##          metro
## married   Outskirts Metro
##   Single         75   344
##   Married       145   436
```

```
chisq.test(metro_married)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  metro_married
## X-squared = 6.6602, df = 1, p-value = 0.009859
```

- **Conclusion:** With $p - value = 0.009859$ which is less than $\alpha = 0.05$, we reject null hypothesis. Thus, there is enough evidence to support the hypothesis that there is some relationship between the marriage and choices to live in metro.

4 Regression

We choose *wage* to be the target variable to build the linear regression model with respect to the remaining quantitative variables *educ*, *exper*, and *hrswk*.

4.1 Simple Linear Regression

4.1.1 Linear regression model of *wage* and *educ*

- We have the linear regression equation is:

$$\widehat{wage} = \beta_0 + \beta_1 \times educ + \epsilon$$

```
educ_wage = lm(wage~educ)
educ_wage
```

```
##
## Call:
## lm(formula = wage ~ educ)
##
## Coefficients:
## (Intercept)         educ
##       -6.71         1.98
```

- Therefore, the linear regression equation is:

$$\widehat{wage} = -6.71 + 1.96 \times educ + \epsilon$$

- Summary of linear model

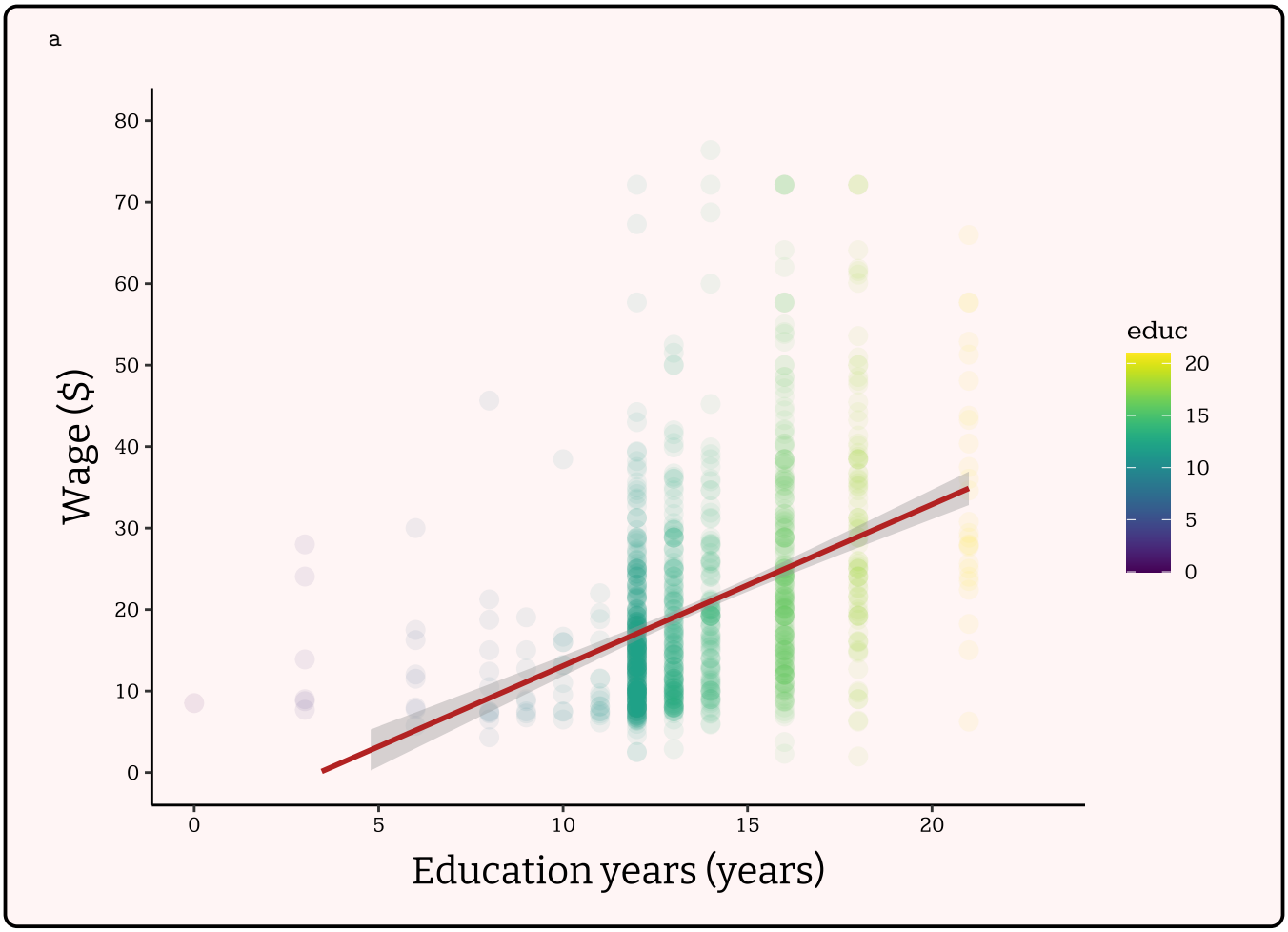
```
summary(educ_wage)
```

```
##
## Call:
## lm(formula = wage ~ educ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.626  -7.816  -2.623   5.019  55.376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.7103     1.9142  -3.506 0.000476 ***
## educ          1.9803     0.1361  14.548  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.66 on 998 degrees of freedom
## Multiple R-squared:  0.175,  Adjusted R-squared:  0.1741
## F-statistic: 211.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

- **Inference**:
  - Coefficients:
    - $\beta_0 = -6.7103$: if we do not know any information about employee's education level, the wage they receive is $-\$6.7103$.
    - $\beta_1 = 1.9803$: if the employees' years of education increase by 1, their wages increase by \$1.9803.
  - Standard Error:
    - $se(b_0) = 1.9142$
    - $se(b_1) = 0.1361$
  - t value:
    - $t - value_0 = -3.506$
    - $t - value_1 = 14.548$
  - Residual Standard Error: measure the difference between predicted wage and the workers' true wage.
    - $\sigma = 11.66$
  - R-squared:
    - $R^2 = 0.1741$: Approximately 17.41% the employees' wage can be accounted for by their education years.
  - p-value:
    - $p - value < 2.2e - 16$: very small compare to significant level $\alpha = 0.05$, hence, the parameter *educ* is different from 0, and *educ* play an important role in determining earnings.
- Plot the linear regression

```
a <- ggplot(mapping=aes(educ, wage)) +
  geom_point(aes(colour = educ), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222") +
  labs(x = "Education years (years)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(0, 80)) +
  xlim(0, 23)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```



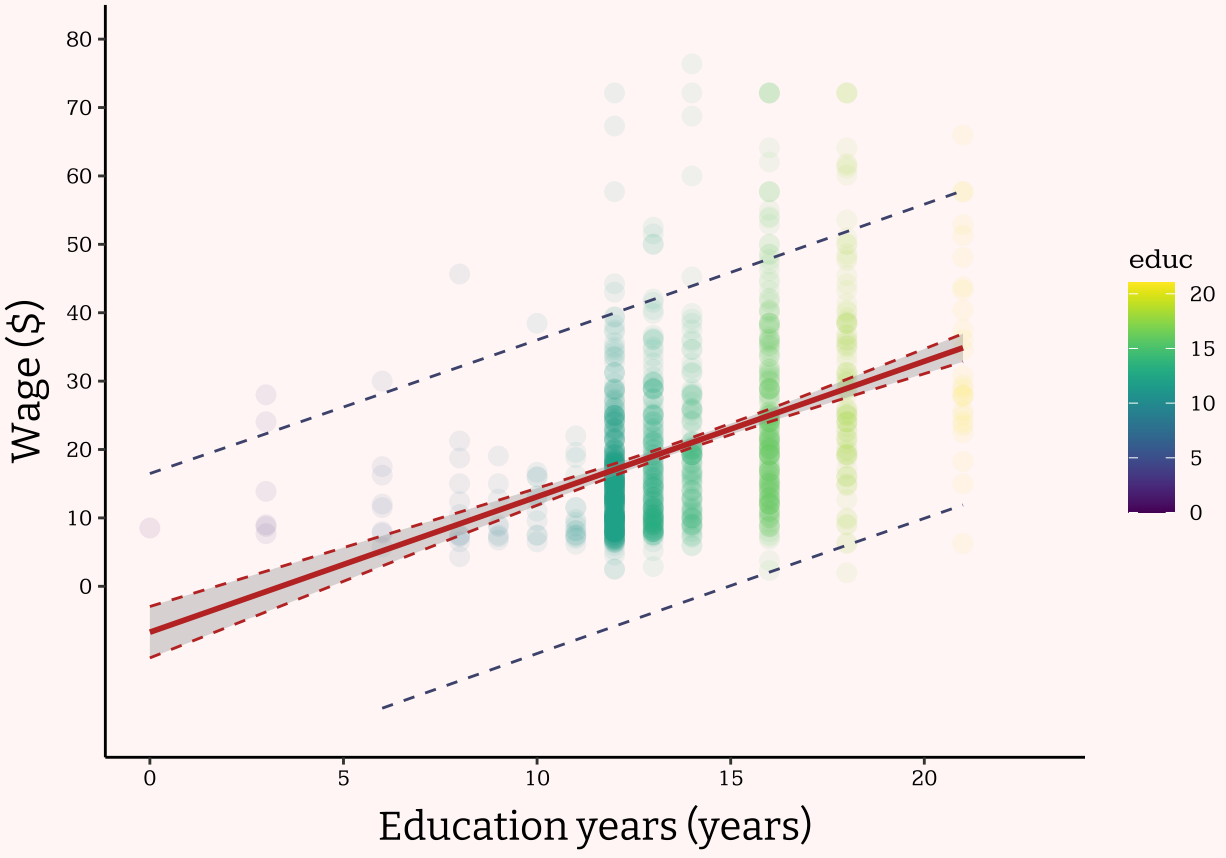
- Estimate 95% confidence interval for the coefficients

```
confint(educ_wage)

##                   2.5 %    97.5 %
## (Intercept) -10.466560 -2.954096
## educ          1.713178  2.247397
```

- **Conclusion:**
  - 95% confidence interval for β_0 (*intercept*) is in range $(-10.466560; -2.954096)$.
  - 95% confidence interval for β_1 (*educ*) is in range $(1.713178; 2.247397)$.
- Plot of linear model with confidence and prediction interval

```
p = predict(educ_wage, interval="prediction")
c = predict(educ_wage, interval="confidence")

colnames(p) = c("p.fit", "p.lwr", "p.upr")
colnames(c) = c("c.fit", "c.lwr", "c.upr")

educ_new <- data.frame(educ, wage, p, c)
```

```
ggplot(educ_new, mapping=aes(x=educ, y=wage)) +
  geom_point(aes(colour = educ), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222") +
  geom_line(aes(y=c.lwr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=c.upr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=p.lwr), colour="#3D426B", linetype="dashed") +
  geom_line(aes(y=p.upr), colour="#3D426B", linetype="dashed") +
  labs(x = "Education years (years)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(-20, 80)) +
  xlim(0, 23)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```



4.1.2 Linear regression model of *wage* and *exper*

- We have the linear regression equation is:

$$\widehat{wage} = \beta_0 + \beta_1 \times exper + \epsilon$$

```
exper_wage = lm(wage~exper)
exper_wage

##
## Call:
## lm(formula = wage ~ exper)
##
## Coefficients:
## (Intercept)        exper
##    18.25768      0.08895
```

- Therefore, the linear regression equation is:

$$\widehat{wage} = 18.25768 + 0.08895 \times exper + \epsilon$$

- Summary of linear model

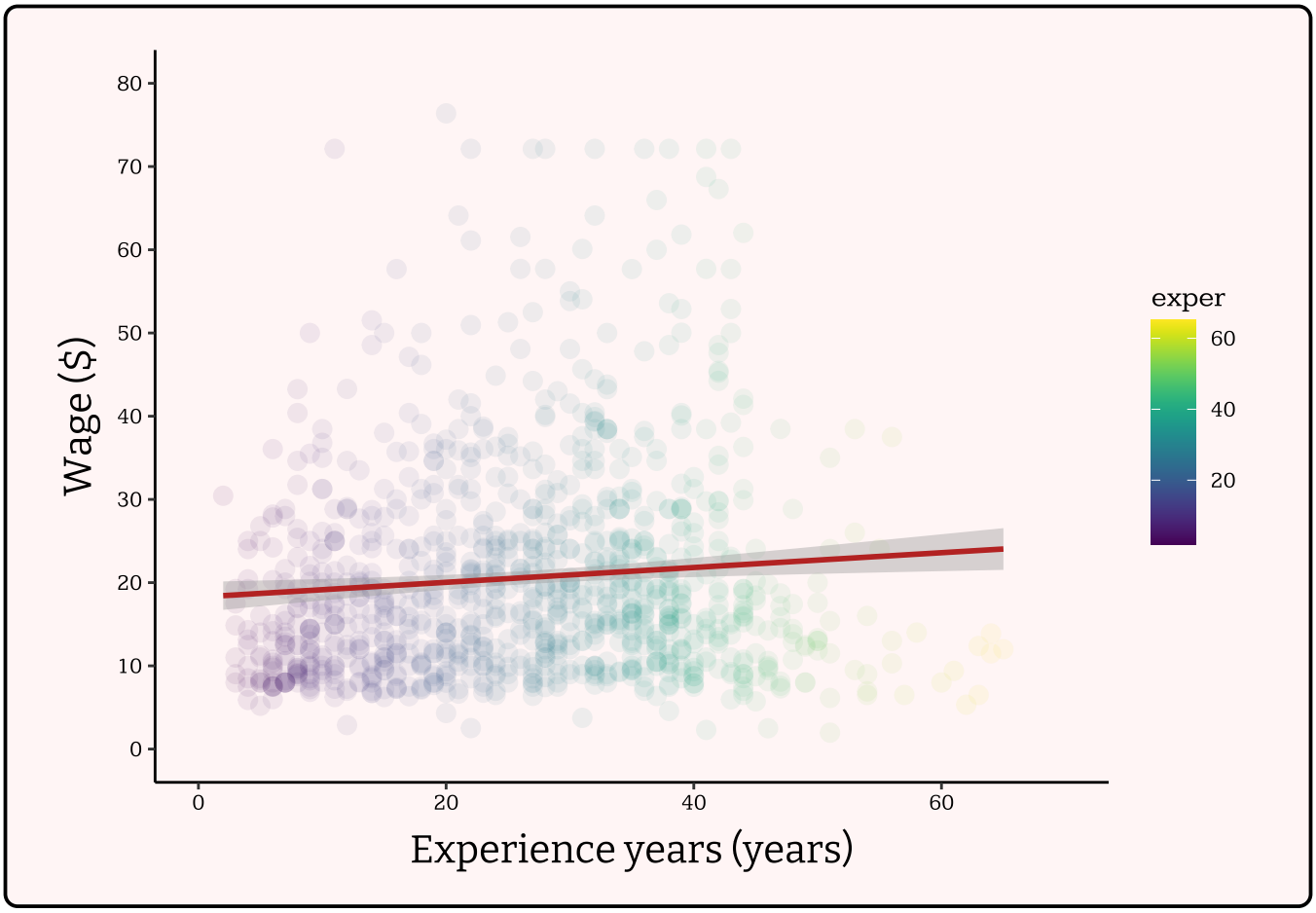
```
summary(exper_wage)

##
## Call:
## lm(formula = wage ~ exper)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.824  -9.224  -3.429   5.422  56.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.25768    0.92733  19.688  < 2e-16 ***
## exper        0.08895    0.03148   2.826  0.00481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.79 on 998 degrees of freedom
## Multiple R-squared:  0.007937,   Adjusted R-squared:  0.006943
## F-statistic: 7.985 on 1 and 998 DF,  p-value: 0.004812
```

- **Inference**:
  - Coefficients:
    - $\beta_0 = 18.25768$: if we do not know any information about employee's experience, the wage they receive is \$18.25768.
    - $\beta_1 = 0.08895$: if the employees' years of experience increase by 1, their weekly earnings increase by \$0.03148.

- Standard Error:
  - $se(b_0) = 0.92733$
  - $se(b_1) = 0.03148$
- t value:
  - $t - value_0 = 19.688$
  - $t - value_1 = 2.826$
- Residual Standard Error: measure the difference between predicted wage and the workers' true wage.
  - $\sigma = 12.79$
- R-squared:
  - $R^2 = 0.006943$: Approximately 0.6943% the employees' wage can be explained by their experience years.
- p-value:
  - $p - value = 0.004812$: smaller than significant level $\alpha = 0.05$, hence, the parameter *exper* is different from 0, and *exper* also contributing in how much the workers are paid per week.

- Plot the linear regression

```
b <- ggplot(mapping=aes(exper, wage)) +
  geom_point(aes(colour = exper), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222", level=0.95) +
  labs(x = "Experience years (years)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(0, 80)) +
  xlim(0, 70)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

b
```



- Estimate 95% confidence interval for the coefficients

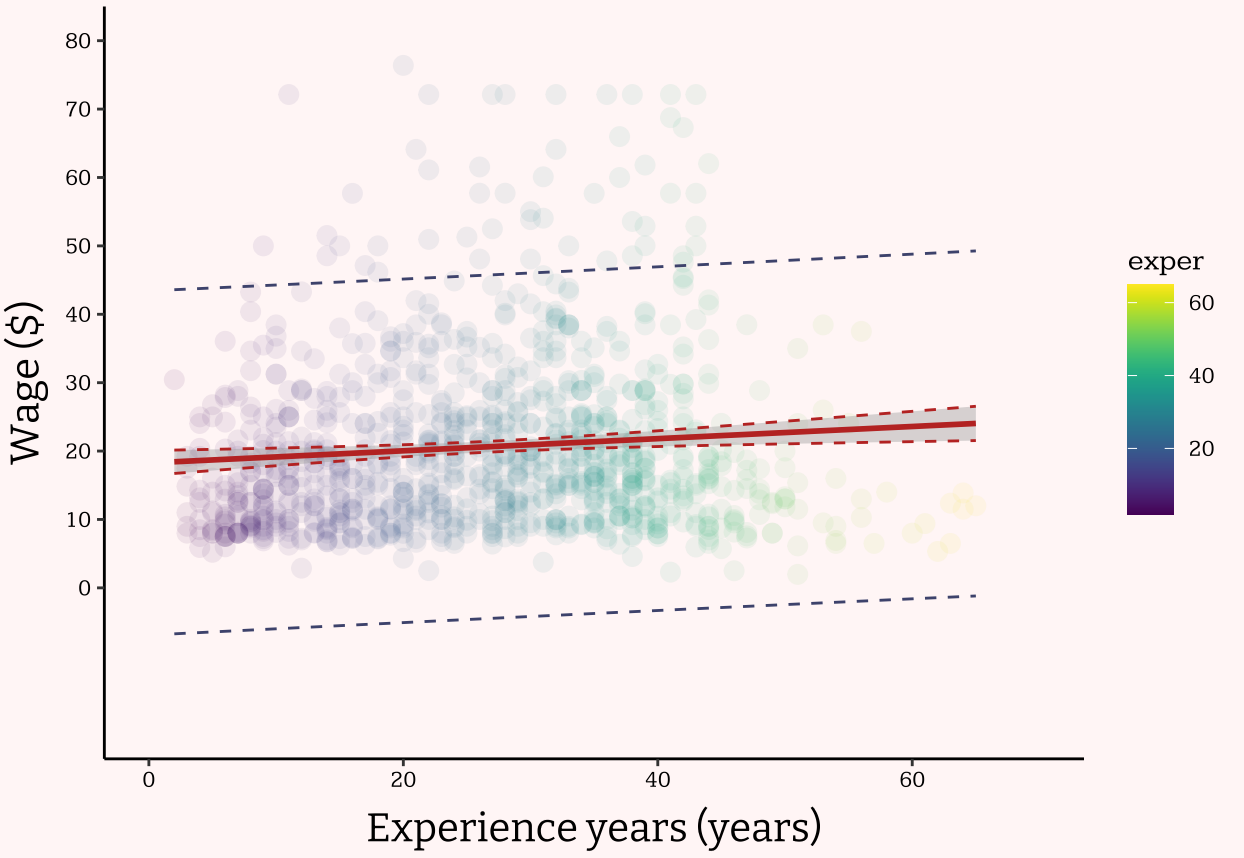
```
confint(exper_wage)
```

```
##                   2.5 %     97.5 %
## (Intercept) 16.43794629 20.0774187
## exper        0.02717852  0.1507283
```

- **Conclusion:**
  - 95% confidence interval for β_0 (*intercept*) is in range $(16.43794629; 20.0774187)$.
  - 95% confidence interval for β_1 (*exper*) is in range $(0.02717852; 0.1507283)$.
- Plot of linear model with confidence and prediction interval

```
p = predict(exper_wage, interval="prediction")
c = predict(exper_wage, interval="confidence")

colnames(p) = c("p.fit", "p.lwr", "p.upr")
colnames(c) = c("c.fit", "c.lwr", "c.upr")

exper_new <- data.frame(exper, wage, p, c)
```

```
ggplot(exper_new, mapping=aes(x=exper, y=wage)) +
  geom_point(aes(colour = exper), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222") +
  geom_line(aes(y=c.lwr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=c.upr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=p.lwr), colour="#3D426B", linetype="dashed") +
  geom_line(aes(y=p.upr), colour="#3D426B", linetype="dashed") +
  labs(x = "Experience years (years)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(-20, 80)) +
  xlim(0, 70)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```



4.1.3 Linear regression model of *wage* and *hrswk*

- We have the linear regression equation is:

$$\widehat{wage} = \beta_0 + \beta_1 \times hrswk + \epsilon$$

```
hrswk_wage = lm(wage~hrswk)
hrswk_wage

##
## Call:
## lm(formula = wage ~ hrswk)
##
## Coefficients:
## (Intercept)        hrswk
##      12.943        0.192
```

- Therefore, the linear regression equation is:

$$\widehat{wage} = 12.943 + 0.192 \times hrswk + \epsilon$$

- Summary of linear model

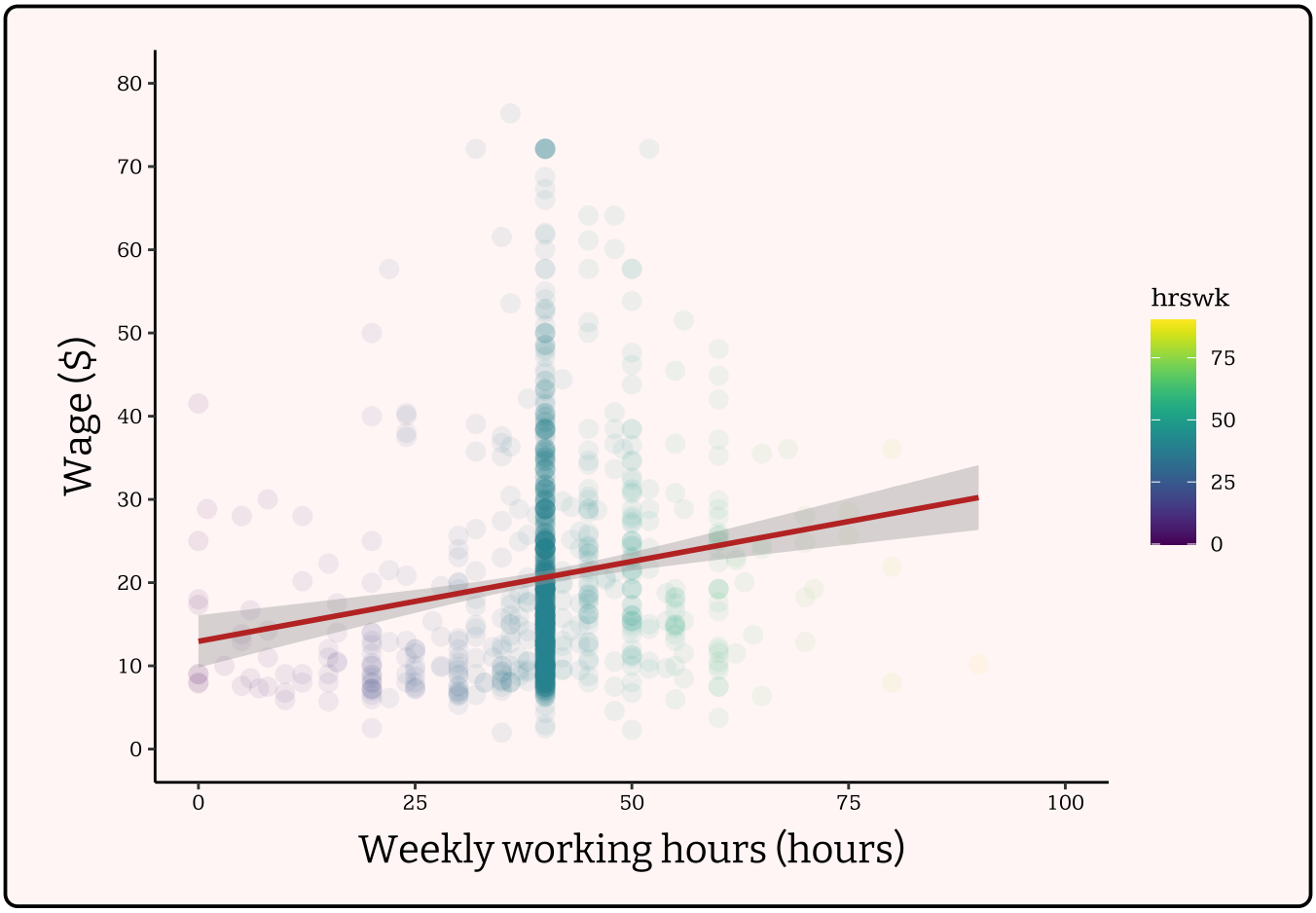
```
summary(hrswk_wage)

##
## Call:
## lm(formula = wage ~ hrswk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.716  -8.626  -3.625   4.703  56.533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.94327    1.60263   8.076 1.91e-15 ***
## hrswk        0.19204    0.03884   4.945 8.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 998 degrees of freedom
## Multiple R-squared:  0.02391,    Adjusted R-squared:  0.02294
## F-statistic: 24.45 on 1 and 998 DF,  p-value: 8.937e-07
```

- **Inference**:
  - Coefficients:
    - $\beta_0 = 12.94327$: if we do not know any information about how many hours the employees work in a week, the wage they receive is \$12.94327.
    - $\beta_1 = 0.19204$: if the employees work 1 more hour, their wages increase by \$0.19204.
  - Standard Error:
    - $se(b_0) = 1.60263$
    - $se(b_1) = 0.03884$

- t value:
  - $t - value_0 = 8.0761$
  - $t - value_1 = 4.945$
- Residual Standard Error: measure the difference between predicted wage and the workers' true wage
  - $\sigma = 12.69$
- R-squared:
  - $R^2 = 0.02294$: Approximately 2.294% the employees' wage can be accounted for by their weekly working hours.
- p-value:
  - $p - value = 8.937e-07$: very small compare to significant level $\alpha = 0.05$, hence, the parameter $hrswk$ is different from 0, and $hrswk$ has an unignorable effect on employees' wages.

- Plot the linear regression

```
c <- ggplot(mapping=aes(hrswk, wage)) +
  geom_point(aes(colour = hrswk), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222", level=0.95) +
  labs(x = "Weekly working hours (hours)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(0, 80)) +
  xlim(0, 100)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))

c
```



- Estimate 95% confidence interval for the coefficients

```
confint(hrswk_wage)
```

```
##                 2.5 %     97.5 %
## (Intercept) 9.7983566 16.0881786
## hrswk       0.1158294  0.2682512
```

- **Conclusion:**
  - 95% confidence interval for β_0 (*intercept*) is in range $(9.7983566; 16.0881786)$.
  - 95% confidence interval for β_1 (*hrswk*) is in range $(0.1158294; 0.2682512)$.
- Plot of linear model with confidence and prediction interval

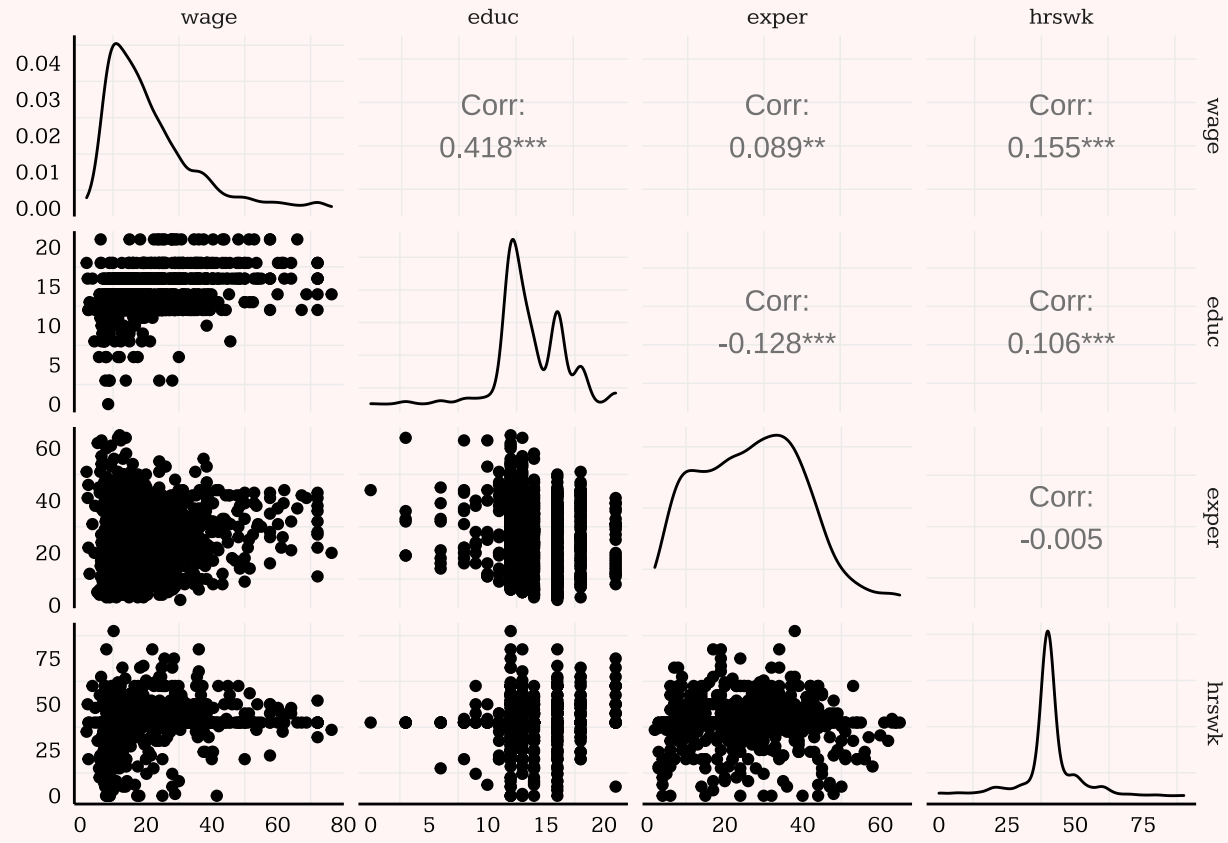
```
p = predict(hrswk_wage, interval="prediction")
c = predict(hrswk_wage, interval="confidence")

colnames(p) = c("p.fit", "p.lwr", "p.upr")
colnames(c) = c("c.fit", "c.lwr", "c.upr")

hrswk_new <- data.frame(hrswk, wage, p, c)
```

```
ggplot(hrswk_new, mapping=aes(x=hrswk, y=wage)) +
  geom_point(aes(colour = hrswk), alpha = 1/12, size = 3) +
  scale_color_viridis() +
  geom_smooth(method='lm', formula= y~x, colour="#B22222") +
  geom_line(aes(y=c.lwr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=c.upr), colour="#B22222", linetype="dashed") +
  geom_line(aes(y=p.lwr), colour="#3D426B", linetype="dashed") +
  geom_line(aes(y=p.upr), colour="#3D426B", linetype="dashed") +
  labs(x = "Weekly working hours (hours)", y="Wage ($)") +
  scale_y_continuous(breaks = seq(0, 80, 10), limits = c(-20, 80)) +
  xlim(0, 100)+
  theme_classic() +
  theme(plot.background = element_blank(),
        panel.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.background = element_blank(),
        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.title.y = element_text(margin=margin(0, 10, 0, 0)),
        axis.title.x = element_text(margin=margin(10, 0, 0, 0)),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```

4.2 Multiple Linear Regression

- The pair plot represent the correlation of each variable and the *wage*

```
df <- data[, c('wage', 'educ', 'exper', 'hrswk')]
ggpairs(df) +
  theme_minimal()+
  theme(panel.grid.major = element_blank(),

        axis.line = element_line(color = "black"),
        axis.title = element_text(family = "Axis", size=15),
        axis.text = element_text(family = "Text", color="#000000"),
        text = element_text(family = "Text", color="#000000"))
```



- We have the linear regression equation is:

$$\widehat{wage} = \beta_0 + \beta_1 \times educ + \beta_2 \times exper + \beta_3 \times hrswk + \epsilon$$

```
mul_wage = lm(wage~educ+exper+hrswk)
mul_wage

##
## Call:
## lm(formula = wage ~ educ + exper + hrswk)
##
## Coefficients:
## (Intercept)         educ        exper        hrswk
##    -16.4432       2.0120       0.1437       0.1373
```

- Therefore, the linear regression equation is:

$$\widehat{wage} = -16.4432 + 2.0120 \times educ + 0.1437 \times exper + 0.1373 \times hrswk + \epsilon$$

- Summary of linear model

```
summary(mul_wage)

##
## Call:
## lm(formula = wage ~ educ + exper + hrswk)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.938  -7.431  -2.565   4.806  56.848
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.44323    2.44008  -6.739 2.70e-11 ***
## educ          2.01199    0.13538  14.861  < 2e-16 ***
## exper         0.14371    0.02839   5.061 4.96e-07 ***
## hrswk         0.13731    0.03522   3.898 0.000103 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.44 on 996 degrees of freedom
## Multiple R-squared:  0.2077, Adjusted R-squared:  0.2053
## F-statistic: 87.03 on 3 and 996 DF,  p-value: < 2.2e-16
```

- **Inference**:
  - Coefficients:
    - $\beta_0 = -16.44323$: if we do not know any information about employee's education level, years of experience and the number of hours they work in a week, the wage they receive is $-\$16.44323$.
    - $\beta_1 = 2.01199$: if the employees' years of education increase by 1, their wages increase by \$2.01199.
    - $\beta_2 = 0.14371$: if the employees' years of experience increase by 1, their wages increase by \$0.14371.
    - $\beta_3 = 0.13731$: if the employees work for 1 more hour, their wages increase by \$0.13731.
  - Standard Error:
    - $se(b_0) = 2.44008$
    - $se(b_1) = 0.13538$
    - $se(b_2) = 0.02839$
    - $se(b_3) = 0.03522$
  - t value:
    - $t - value_0 = -6.739$
    - $t - value_1 = 14.861$
    - $t - value_2 = 5.061$
    - $t - value_3 = 3.898$
  - Residual Standard Error: measure the difference between predicted wage and the workers' true wage.
    - $\sigma = 11.44$

- R-squared:
  - $R^2 = 0.2053$: Approximately 20.53% the employees' wage can be explained by their years of education, years of experience and weekly working hours.
- p-value:
  - $p - value < 2.2e - 16$: very small compare to significant level $\alpha = 0.05$, hence, the parameters *educ*, *exper*, and *hrswk* are all different from 0. By seeing the mark *** representing the significant code in three parameters, we can conclude that all of them have remarkable impact on employees' wages.

- Estimate 95% confidence interval for the coefficients

```
confint(mul_wage)

##                     2.5 %      97.5 %
## (Intercept) -21.23151578 -11.6549346
## educ          1.74632424   2.2776647
## exper         0.08798955   0.1994280
## hrswk         0.06819347   0.2064331
```

- **Conclusion:**
  - 95% confidence interval for β_0 ($intercept$) is in range $(-21.23151578; -11.6549346)$.
  - 95% confidence interval for β_1 ($educ$) is in range $(1.74632424; 2.2776647)$.
  - 95% confidence interval for β_2 ($exper$) is in range $(0.08798955; 0.1994280)$.
  - 95% confidence interval for β_3 ($hrswk$) is in range $(0.06819347; 0.2064331)$.

4.3 Conclusion

As the results presented above, all three attributes *educ*, *exper*, and *hrswk* contribute in deciding how much the workers earned. Based on Residual Standard Error implicating the difference between predicted value and the true one, σ of multiple linear regression (11.44) is the smallest comparing to other simple linear model, so the multiple linear model predict the wage more accurately. Moreover, the index of $R - squared$ of the multiple linear regression with the present of all three variables explains the highest proportion of the wage, which is 20.53%. Therefore, we should choose the multiple linear regression that has the equation:

$$\widehat{wage} = -16.4432 + 2.0120 \times educ + 0.1437 \times exper + 0.1373 \times hrswk + \epsilon$$

as the best model to predict the employees' wage.

5 Summary

- By data visualization:
  - The number of workers receive the high wage decreases as the wage increases.
  - The two most popular education years workers have are 12 and 16, which represents High school graduation and Bachelor's degree.
  - The experience years of employees varies a lot, mostly from 9 to 39 years.
  - Approximately 70% of the employees in the survey work 40 hours per week.
  - The rate of female and male workers in the data is almost the same.
  - The number of people living the metro is significantly higher than other places.

  - The number of employee living in area four areas (Mid West, South, West, and Others) is almost equal to each other.
  - The rate of Asian and Black people in the survey is not too many, which are outweighed by other races.

- By testing:
  - The mean wage of male is greater than female.
  - There is not enough evidence to state that the single employee work more hours than the married one.
  - The male's rate of marriage is higher than that of female.
  - The proportion of Asian living in Metropolitan is bigger than other races.
  - There is not sufficient evidence to support the claim that marriage relates to gender.
  - There is some relationship between marriage and the decision to live in big city.

- Regression:
  - All three variables *educ*, *exper* and *hrswk* affect the wage the worker earned.
  - The most effective linear regression model has equation:

$$\widehat{wage} = -16.4432 + 2.0120 \times educ + 0.1437 \times exper + 0.1373 \times hrswk + \epsilon$$